Spring 1997

# Judging contingencies accurately: The effects of feedback, practice, and self-efficacy

Steven C. Clark
*University of New Hampshire, Durham*

# INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

**The quality of this reproduction is dependent upon the quality of the copy submitted.** Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

**UMI**

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor MI 48106-1346 USA
313/761-4700 800/521-0600

# JUDGING CONTINGENCIES ACCURATELY: THE EFFECTS OF FEEDBACK, PRACTICE, AND SELF-EFFICACY

BY

STEVEN C. CLARK

B.S. Brigham Young University, 1992
M.A. University of New Hampshire, 1994

DISSERTATION

Submitted to the University of New Hampshire
in Partial Fulfillment of the Requirements
for the Degree of

Doctor of Philosophy

in

Psychology

May, 1997

This dissertation has been examined and approved.

Victor A Benassi

Dissertation Director, Victor A. Benassi
Professor of Psychology

Rebecca M. Warner

Rebecca M. Warner
Professor of Psychology

John E. Limber

John E. Limber
Associate Professor of Psychology

Gary S. Goldstein
University of New Hampshire at Manchester
Associate Professor of Psychology

Susan E. Manfull

Susan E. Manfull
University of New Hampshire
Instructor

5/1/97

Date

# DEDICATION

To my parents,

Wayne and Colleen Clark

And to my advisor,

Victor A. Benassi

ACKNOWLEDGMENTS

Over the past five years, Victor Benassi has been a friend and a mentor. He has always been supportive and understanding.

Becky Warner has provided support and valuable suggestions on both a Master's Thesis and a Doctoral Dissertation. John Limber was a great help with the Hypercard program that allowed me to conduct these experiments. Gary Goldstein and Susan Manfull have both provided valuable suggestions for the presentation of this research.

Glen Bailey, Lynsey Maxfield, and Amy Adams helped with the collection of data and kept me company during long days in the lab.

Dr Pepper® kept me going during difficult times. I wouldn't have finished on time without it.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

JUDGING CONTINGENCIES ACCURATELY:
THE EFFECTS OF FEEDBACK, PRACTICE, AND SELF-EFFICACY

by

Steven C. Clark
University of New Hampshire, May, 1997

Some psychologists have claimed that people are not good at judging covariation (e.g., Smedslund, 1963; Jenkins & Ward, 1965). This claim, however, has been based on the results of experiments that may not have been optimal for promoting judgmental accuracy (Allan & Jenkins, 1980). Other psychologists have claimed that people are relatively good judges of covariation (e.g., Wasserman, Chatlosh, & Neunaber, 1983; Wasserman & Shaklee, 1984). Common to most of this research is an experimental paradigm in which participants do not ever receive feedback concerning the accuracy of their judgments.

The two experiments in this dissertation were designed to promote accuracy in the judgment of contingency by providing (a) accurate feedback concerning participants' judgments and (b) practice with judging many contingency problems. The results of these experiments indicate that people become better judges of contingency with feedback and

practice, but do not improve with practice alone. This is true for positive contingencies (Experiment 1) and negative contingencies (Experiment 2). Judgmental accuracy was greatest for extreme contingency problems ($\Delta P$ less than -.75 and greater than .75).

Self-efficacy has been shown to account for performance in a variety of domains above and beyond ability. Experiment 1 addressed the relation between self-efficacy, feedback, and judgmental accuracy. Mean self-efficacy increased over the course of the experiment for participants in the feedback condition, but decreased for participants in the no feedback condition. Participants with high self-efficacy in the feedback condition showed relatively accurate judgments of contingency, but participants with high self-efficacy in the no feedback condition showed relatively inaccurate judgments of contingency.

Experiment 1 also addressed whether judgmental accuracy on one contingency task transferred to judgmental accuracy on a different task. The results indicate that there was no transfer in accuracy from one task to another.

# INTRODUCTION

An important prerequisite of adaptive behavior is sensitivity to covariation. Without some sense of what events are related, adaptive behavior can occur only by accident. And having occurred by accident, adaptive behavior will continue as a matter of chance if one is not sensitive to the relation between responses and outcomes (Herrnstein, 1966). Perhaps for this reason, considerable research has focused on people's ability to judge the covariation between events. (See Alloy & Tabachnik, 1984; Crocker, 1981; Shaklee, 1983; and Wasserman, 1990) Studied in different ways and for different purposes, this research has been known variously as the judgment of covariation, correlation, contingency, control, and causation. (Covariation is the most general of the terms and does not denote a causal relation between the variables. Contingency, the term I use to refer to the present research, does denote a causal relation between the variables.) The judgment of covariation has been of interest to psychologists because it addresses people's sensitivity to the relation between events.

Smedslund (1963) asked nursing students to judge the correlation between variables on a number of different tasks. In one task, the students sorted through a deck of 100 cards which each contained eight letters: four of them representing a symptom (chosen from among the letters A, B, C, D, or E)

and four of them representing a diagnosis (chosen from among the letters F, G, H, I, or J). Participants judged the relation between symptom A and diagnosis F based on what they observed in the deck of cards. Smedslund's results indicated that his participants were not good judges of the empirical correlation, leading him to conclude that

> normal adults with no training in statistics do not have a cognitive structure isomorphic with the concept of correlation. Their strategies and inferences reveal a particularistic, non-statistical approach, or an exclusive dependence on the frequency of [positive confirming] instances. (p. 172)

Other psychologists have found that people can make relatively accurate covariation judgments. Wasserman and Shaklee (1984) presented college students with a scenario in which a person (Kim) was trying to fix a malfunctioning radio. Kim's tapping on one of the radio's internal wires and the radio's intermittent buzzing were presented on time lines which recorded when each occurred. Wasserman and Shaklee found that participants accurately judged the contingency problems.

While Smedslund (1963) and Wasserman and Shaklee (1984) came to different conclusions about people's ability to judge covariation, their experiments were procedurally similar in one important way--participants were never informed of their degree of judgmental accuracy. This lack of feedback is common to most judgment of covariation research (Hogarth,

1981). Unfortunately, this creates an experimental setting quite different from the real-world in which we typically receive feedback (McFarland, 1971). Further, many experiments require participants to make only a few judgments. In the real-world, we make many judgments as part of an ongoing process. Most judgment of covariation experiments, however, lack these elements.

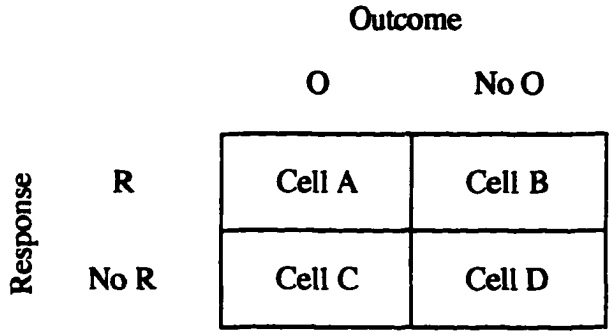
My primary purpose in the present research was to evaluate judgmental accuracy with and without feedback. Participants judged a number of problems during the experimental session to assess whether practice would improve judgmental accuracy. I also investigated the role of self-efficacy in the judgment of contingency and whether there is a transfer of judgmental accuracy from one judgment task to another.

A Note on Variables and Measures of Covariation

Some covariation research has utilized more than two variables or variables that can have more than two states. The majority of the research in this area, however, has been done with two binary variables (Allan, 1993). When research is done with two binary variables, all possible combinations of the two variables can be recorded in a 2 X 2 table (see Figure 1). On the top of the table are the two states of one variable, such as the illumination of a light occurring (O) or not occurring (no O). On the side of the table are the two

states of the other variable, such as pressing a telegraph key (R) or not (no R).

Figure 1
The Cells of a 2 X 2 Table Representing the Response-Outcome Possibilities




Instances of Cell A have been called "positive confirming" cases because both variables are present or co-occur (R, O). Instances of Cells B and C have been called "disconfirming cases" because one variable is present but the other is not (R, no O and no R, O). Instances of Cell D have been called "negative confirming" cases because neither variable is present (no R, no O).

The example just given can be classified as a one-response one-outcome contingency task (Allan, 1993). In a one-response one-outcome contingency task (1R/1O), participants can make a single active response on each trial (e.g., press a button or do not press it), and then a single active outcome may result (e.g., a light is illuminated or it is not illuminated). In a two-response two-outcome contingency task (2R/2O), participants can choose which of

two active responses to make on each trial (e.g., press button A or press button B), and one of two active outcomes occurs (e.g., light Y is illuminated or light Z is illuminated). The difference between a 1R/10 task and a 2R/20 task may seem trivial, but Allan & Jenkins (1980) found that participants made more accurate judgments on a 1R/10 task than a 2R/20 task. In addition, the above nomenclature of confirming and disconfirming cases loses its relevance in a 2R/20 task. Instead, there are simply four types of events. The experiments conducted for this dissertation employ 1R/10 tasks.

The actual relation between two binary variables can be statistically defined in a number of ways (Allan, 1980). The most common measure in the judgment of contingency literature is $\Delta P$, the probability of an outcome given a response minus the probability of an outcome given no response ($\Delta P$ = p[O/R] - p[O/no R] or in terms of the four cells of the 2 X 2 table, $\Delta P$ = A/[A+B] - C/[C+D]). $\Delta P$ reflects a one-way contingency or relation, such as the relation between one's pressing a key and the illumination of a light. $\Delta P$ values can range from -1.00 to 1.00. $\Delta P$ = 1.00 indicates a perfect positive contingency (e.g., a press always causes a flash, a flash never occurs without a press). $\Delta P$ = -1.00 indicates a perfect negative contingency (e.g., a press always prevents a flash, a flash always occurs without a press). $\Delta P$ = .00 indicates

noncontingency (e.g., pressing has no effect on the flashing).

There are other measures of covariation which reflect a joint relation between the variables. When both variables influence the outcome of the other, $\chi^2$ and $\phi$ are often used to statistically define the relation ($\chi^2 = N[AD-BC]^2 / [(A+B)(C+D)(A+C)(B+D)]$ and $\phi$ = square root of $\chi^2/N$). These measures are less common than $\Delta P$ in the judgment of contingency literature (Allan, 1980).

## Research Indicating Judgmental Inaccuracy

### Smedslund (1963)

One of the early investigations of the judgment of correlation was conducted by Smedslund (1963). He presented several paper-and-pencil judgment of contingency tasks to student nurses. The first was the card sorting task that was described above. This was a serial presentation of the data, meaning that the data were presented sequentially. In a second task, participants judged the relation between symptom A and diagnosis F based on summary information which reported the cell frequencies for a 2 X 2 table. In a third task, students sorted through a deck of 100 cards which contained information only about the presence or absence of symptom A (+A or -A) and diagnosis F (+F or -F). Regardless of the task they were given, students' judgments were unrelated to the actual contingency. Smedslund reported that to the extent

that his participants reasoned statistically, they relied almost exclusively on Cell A cases.

Jenkins and Ward (1965)

Jenkins and Ward (1965) also reported that people are not good judges of contingency. Participants in Jenkins and Ward's experiments participated in a two-response two-outcome contingency task (2R/2O). Participants could press either of two buttons and then observe the illumination of one of two lights. The contingency problems consisted of 60 self-paced discrete trials. In a discrete-trial task there are clearly defined trials which each have a period for a response and an outcome. In Jenkins & Ward's experiments, participants pressed one of the two buttons and then pressed a "test" button which illuminated one of the two outcome lights. After one of the lights had been illuminated for two seconds, the experimental apparatus was automatically reset for the next trial. Participants completed five contingency problems.

In their first experiment, Jenkins and Ward told some participants that their task was to "score" as many points as possible by causing a designated light to illuminate as a result of their button pressing. Jenkins and Ward told other participants that their task was to "control" the illumination of the two lights through their button pressing. Yoked to these active participants, observer participants watched the responses and outcomes on a display panel in another room. Regardless of participants' experimental

condition ("score" or "control" instructions, active or observer role), Jenkins and Ward found that there was little relation between participants' judgments and the actual contingency of the problems they completed.

In Jenkins and Ward's (1965) second experiment, participants judged the same contingency problems as in Experiment 1 and answered additional questions concerning their ability to control the illumination of the lights. They also answered questions which addressed their understanding of the concepts of probability. Again, participants' judgments were unrelated to the actual contingency of the problems.

In Jenkins and Ward's third experiment, participants completed two training problems before completing the same problems that were used in the previous experiments. One of the training problems had $\Delta P = .00$ and the other had $\Delta P = .80$. Even after these training problems, participants' judgments bore little relation to the actual contingencies of the test problems.

## Research Indicating Judgmental Accuracy

Smedslund (1963) and Jenkins and Ward (1965) concluded that people are poor judges of covariation. They based their conclusion on what most psychologists would regard as poor performance on the part of their participants. In situations like these, it is easy to say how people fared. But at what point does one conclude that people are good judges of

covariation? In other words, what are the criteria of judgmental accuracy? Even in the articles that suggest that people are good judges of contingency, there are no universal criteria for judgmental accuracy. But there are indications that participants can make relatively accurate judgments.

Wasserman, Chatlosh, and Neunaber (1983)

Wasserman, Chatlosh, and Neunaber (1983) found that participants made accurate contingency judgments on a free-operant 1R/10 task. A free-operant task is procedurally different from a discrete-trial task in that there is not any sort of inter-trial interval or marker. Instead, participants can respond at any time during the problem. Outcomes are presented at the end of experimentally defined sampling intervals, such as one second.

In their experiments, Wasserman et al. (1983) had participants judge the effect of their key pressing on the illumination of a light, recording their judgments on a 201 point scale (-100 = prevents the light from occurring; 100 = causes the light to occur). Participants could respond or not respond at any time during the experiment and were instructed to observe what happened when they did and did not respond. Each participant completed nine contingency problems in a randomly determined order. The probability of an outcome given a response (p[O/R]) and the probability of an outcome given no response (p[O/no R]) were one of three levels, .125, .500, and .875. In combination, these probabilities produced

nine problems with five levels of contingency: $\Delta P$ = -.750, -.375, .000, .375, and .750.

Wasserman et al. (1983, Experiment 1) instructed participants in one condition to respond by tapping the telegraph key but not hold it down (tap condition) and in another condition to respond by pressing the key down and holding it (press condition). Each problem consisted of 240 one-second sampling intervals. Wasserman et al. found relatively accurate judgments in both experimental conditions. For example, participants in the tap condition gave the following ratings for the five levels of contingency: $\Delta P$ = -.750, $M$ = -68; $\Delta P$ = -.375, $M$ = -30; $\Delta P$ = .000, $M$ = 0; $\Delta P$ = .375, $M$ = 30; $\Delta P$ = .750, $M$ = 75.

In their second experiment, Wasserman et al. (1983) manipulated both the length of the sampling interval and the number of sampling intervals. In one condition, participants had 240 one-second sampling intervals per problem. In a second condition, participants had 60 one-second sampling intervals. And in a third condition, participants had 60 four-second sampling intervals. Participants in all three conditions made accurate judgments of contingency.

In their third experiment, Wasserman et al. (1983) manipulated the nature of the sampling interval. In one condition, the sampling interval was a fixed three seconds. In the other condition, the sampling interval was three seconds on average, but any given interval could be one,

three, or five seconds long. Again, participants provided judgments that were sensitive to the actual level of contingency in the problems.

Wasserman and Shaklee (1984)

Wasserman and Shaklee (1984) explored the effects of presenting contingency information in several different paper-and-pencil formats. Some prior research indicated that judgments of contingency are more accurate when the data are presented in a summary table than when they are presented serially (e.g., sorting through a stack of cards). Wasserman and Shaklee proposed that this difference in accuracy may be due to the added memory demands of the serial presentation. When participants see data presented serially, they must not only make a judgment, but they must also try to count and then recall the frequencies of the events. Wasserman and Shaklee designed their experiments to compare summary table presentation and serial presentation in a setting in which the two tasks had similar memory demands. They used time lines which recorded on a single page all of the data for an entire problem.

Wasserman and Shaklee (1984) conducted four experiments in which participants judged 24 contingency problems, recording their judgments on a nine point scale (-4 = prevents the sound from occurring; 0 = has no effect; 4 = causes the sound to occur). The probability of an outcome given a response (p[O/R]) and the probability of an outcome

given no response (p[O/no R]) were one of five levels: .00, .25, .50, .75, and 1.00. In combination, these probabilities produced 24 problems with nine levels of contingency: $\Delta P$ = -1.00, -.75, -.50, -.25, .00, .25, .50, .75, and 1.00.

Participants' judgments of these contingencies were sensitive to the varying levels of $\Delta P$ . For example, in Wasserman and Shaklee's (1984) Experiment 2, participants made half of their judgments based on summary tables and half of their judgments based on time lines. Wasserman and Shaklee found that judgments were similar for the two formats. In both formats, the mean judgments of the participants scaled the contingency problems. That is, if the task had been to rank order the contingency problems, the participants' mean judgments would have put them in the correct order, from smallest to largest. For example, judgments for the summary table information were as follows: $\Delta P$ = -1.00, $M$ = -1.44; $\Delta P$ = -.75, $M$ = -1.36; $\Delta P$ = -.50, $M$ = -.71; $\Delta P$ = -.25, $M$ = -.53; $\Delta P$ = .00, $M$ = .22; $\Delta P$ = .25, $M$ = .51; $\Delta P$ = .50, $M$ = 1.17; $\Delta P$ = .75, $M$ = 1.25; $\Delta P$ = 1.00, $M$ = 2.44. Wasserman and Shaklee (1984) found a similar pattern of results in all four of their experiments.

## Attempts to Improve Accuracy Through Training

In contrast to Smedslund (1963) and Jenkins and Ward (1965), Wasserman et al. (1983) and Wasserman and Shaklee (1984) found that people make accurate judgments of contingency. Other experiments have found similar results

(e.g., Allan & Jenkins, 1980, 1983; Alloy & Abramson, 1979; Peterson, 1980; and Ward & Jenkins, 1965). There is little evidence, however, indicating whether people's judgments become more accurate with experience and training. This is the principle focus of the present research. To date, only two experiments have systematically tried to improve judgmental accuracy through training.

Jenkins and Ward (1965)

One attempt to improve the accuracy of participants' judgments was undertaken by Jenkins and Ward (1965). As was mentioned above, Jenkins and Ward found in their Experiments 1 and 2 that participants' judgments of contingency were not related to the actual contingency of the problems. In an attempt to improve participants' accuracy in their third experiment, Jenkins and Ward's (1965) participants completed two training problems prior to performing the same judgment of contingency problems as in Experiments 1 and 2. One of the training problems had $\Delta P = .80$. Prior to starting this problem, Jenkins and Ward told participants, "You will have very good control over the outcomes by your choice of responses" (p. 13). Participants were shown a sample answer sheet that had been marked at 80 (0 = No Control; 100 = Complete Control) and then they completed the problem. The other training problem had $\Delta P = .00$. Prior to starting this problem, Jenkins and Ward told participants, "Your choice of responses will have no influence over which outcome will

appear" (p. 13). Participants were shown a sample answer sheet that had been marked at 0 (0 = No Control; 100 = Complete Control) and then they completed the problem.

As in their Experiments 1 and 2, Jenkins and Ward (1965) found that there was little relation between participant's judgments and the actual contingency of the problem.

One of the reasons why their training may not have been effective is that Jenkins and Ward's experimental task was not conducive to accurate judgments. Jenkins and Ward used a 2R/20 task. Allan and Jenkins (1980) showed in a series of experiments that participants who have a single response option (1R--move a joystick or not move a joystick) provide more accurate judgments than participants who have two response options (2R--move the joystick to the right or move the joystick to the left).

Another explanation for why Jenkins and Ward's (1965) training did not improve judgmental accuracy is that their training was limited to two problems. This is a limited amount of exposure to a relatively complex task.

Clark and Benassi (in press)

Clark and Benassi (in press) also provided some training to participants in a judgment of contingency task. Their experiment was designed to examine Sherif's theory of contrast and assimilation (Sherif, Taub, & Hovland, 1958) in the judgment of contingency. Because Clark and Benassi's interest was in judgmental displacement away from anchoring

judgments, they wanted to establish that participants' initial judgments were relatively accurate.

Clark and Benassi (in press) gave participants four contingency time lines modeled after those of Wasserman and Shaklee (1984) and Newman and Benassi (1989). The first two time lines provided an anchor for participants' judgments of the third and fourth time lines. The initial time lines had $\Delta P$ = .00, .50, or 1.00, depending on experimental condition. The page containing the first time line also had a sentence indicating to participants the actual value of that time line on the judgment scale (0, 5, or 10, respectively, on a 21 point scale: -10 = prevents buzzing, 0 = has no effect, 10 = causes buzzing). The second time line in each condition was identical to the first, but the second page did not contain any information telling the participants the value of that time line on the judgment scale.

How much effect did exposure to an accurate assessment of contingency have on participants? The best way to answer this question is to compare the judgments of participants in Clark and Benassi's (in press) experiment to the judgments of participants in a similar experiment. Newman and Benassi (1989, Experiment 3) also had three experimental conditions with initial time lines of $\Delta P$ = .00, .50, or 1.00. Their participants, however, were not given any training in how to judge the time lines. Newman and Benassi (1989) obtained mean contingency judgments of 1.58, 4.40, and 8.85, respectively,

for the three initial time lines. Clark and Benassi (in press) obtained mean contingency judgments of .89, 4.89, and 9.32, respectively. The mean judgments in Clark and Benassi's experiment were closer to the nominal value of each contingency (.00, .50, and 1.00). This finding appears to be due to the training and to the fact that the first two time lines were identical. This comparison indicates that participants in Clark and Benassi's experiment learned from the first time line. Unfortunately, this experiment does not provide any data concerning long-term improvements in judgmental accuracy or how practice might affect judgments.

In the experiments of Jenkins and Ward (1965) and Clark and Benassi (in press), the amount of training and the total exposure to judging contingencies was minimal. Participants in Jenkins and Ward's Experiment 3 completed a total of seven contingency problems (only two were training problems). Participants in Clark and Benassi's experiment judged four contingency problems (only one was a training problem). The present experiments were designed to overcome these limitations by providing ongoing training (i.e., accurate feedback) to participants for a large number of contingency problems.

## The Discriminability of Covariation

There is evidence from a number of experiments that the relation between objective and judged covariation may not be linear. For example, Well, Boyce, Morris, Shinjo, and

Chumbley (1988) presented participants with three sets of 60 paired numbers. The numbers in each set had a different objective correlation: .10, .60, or .90. After each set was presented to participants, they recorded their judgment on a scale that ranged from 0 (no relationship) to 100 (perfect relationship). The mean judgments of participants who observed the number sets were as follows: $r$ =.10, $M$ = 30.35; $r$ =.60, $M$ = 32.95; $r$ =.90, $M$ = 57.05. The difference between mean judgments of the .10 and .60 correlations was 2.6 units, while the difference between mean judgments of the .60 and .90 correlations was 24.10 units. Even though there was less absolute difference between the .60 and .90 correlations, participants showed more discrimination between them than between the .10 and .60 correlations.

Bobko and Karren (1979) presented participants with scatterplots representing correlations of .00, .35, and .64. Participants estimated the correlation coefficient between the x and y variables to two decimal places. The difference between the median judgments of the .00 (median = .00) and .35 (median = .10) correlation scatterplots was only .10 units. The difference between the median judgments of the .35 and .64 (median = 50) correlation scatterplots was .40 units. Again, there is evidence of a difference in discriminability along the covariation continuum.

Clark and Benassi (in press) noted this difference in discriminability and discussed its implications for

experiments designed to produce context effects (i.e., contrast and assimilation). As an experimenter, one must be mindful that any "midrange" stimuli are midrange not only in objective terms, but also in psychophysical terms.

The above examples (Well et al., 1988, and Bobko & Karren, 1979) and the discussion of Clark and Benassi (in press) focus on the discriminability of covariation between .00 and 1.00. What evidence is there concerning the discriminability of covariation along the continuum from -1.00 to 1.00? Wasserman & Shaklee (1984) reported data that are informative on this topic. In their four experiments, they presented contingencies ranging from -1.00 to 1.00 and found a difference in discriminability between negative and positive contingencies. In their Experiment 3, for example, participants examined time lines which presented Kim's tapping on the radio's wire and the radio's buzzing. Participants then recorded their judgments on a scale that ranged from -4 (prevents sound from occurring) to 4 (causes sound to occur).

Participants in the broken time line condition (a paper-and-pencil equivalent of a discrete trial procedure), showed a difference in discriminability between negative contingencies and positive contingencies. On problems with an outcome probability of .50, the following mean judgments were obtained: $\Delta P$ = -1.00, $M$ = -2.12; $\Delta P$ = -.50, $M$ = .16; $\Delta P$ = .00, $M$ = .36; $\Delta P$ = .50, $M$ = 1.60; $\Delta P$ = 1.00, $M$ = 3.80. The

difference between judgments of the -1.00 and -.50 time lines was 2.28 units. The difference between the -.50 and .00 time lines was .20 units. The difference between the .00 and .50 time lines was 1.24 units. And the difference between the .50 and 1.00 time lines was 2.20 units.

This pattern of mean judgments suggests two conclusions. First, there is less discrimination between problems with low levels of contingency (-.50 to .00) than there is between problems with high levels of contingency (-1.00 to -.50). This is the same pattern of discriminability that has been noted for positive contingencies. Second, there is less discrimination between perfect negative contingency and noncontingency (2.48 units) than there is between perfect positive contingency and noncontingency (3.44 units). This pattern of results indicates that negative contingencies may be more difficult to discriminate among than positive contingencies.

The data of Wasserman and Shaklee (1984) suggest that people are not as good at discriminating among negative contingencies as they are at discriminating among positive contingencies. Their results indicate that the psychophysical function for the judgment of contingency is a gently curving backwards "s" that underestimates the actual contingency of the problems. For this reason, the first experiment of this dissertation was focused on those contingencies that are more discriminable, specifically, contingencies between .00 and

1.00. The second experiment was conducted to examine whether the same pattern of judgmental accuracy would be found for a more difficult to discriminate set of contingencies (contingencies between -1.00 and .00).

## Response Rate and Judgmental Accuracy

When judgmental accuracy is the goal, the best response strategy is to respond on half of the trials. This is the case because one would want to have an equally large number of trials with a response and without a response on which to base one's judgment. Any deviation from responding half of the time reduces the amount of information for either p(O/R) or p(O/no R).

Wasserman et al. (1983) found in their Experiment 1 that there was a difference in judgmental accuracy as a result of the rate of response. Wasserman et al. rank ordered their participants in both conditions (press button or tap button) according to participants' mean probability of response over all the problems. They then performed a median split in each condition. The resulting four groups were press-low, tap-low, press-high, tap-high, with mean response probabilities of .17, .23, .37, and .44, respectively.

The press-high ($\underline{p}$ = .37) and tap-high ($\underline{p}$ =.44) groups had mean judgments of contingency that were very close to the nominal values of the contingencies. The tap-low group ($\underline{p}$ = .23) had mean judgments that were relatively accurate for positive contingencies, but participants in this group

underestimated the degree of relation for negative contingencies. The press-low group ($\underline{p}$ = .17) produced a flattened function in which participants underestimated the actual relation of both negative and positive contingencies.

In the present experiments, participants will be encouraged to respond on about half of the trials to promote judgmental accuracy. Research by Benassi and Mahler (1985) found that participants do respond about half the time when they are instructed to do so. If participants' rate of response, however, is as extreme as that of Wasserman et al.'s (1983) tap-low or press-low groups, their judgmental accuracy should diminish.

## Self-Efficacy

Self-efficacy (Bandura, 1986) refers to the extent to which people are confident they possess the ability to successfully perform certain behaviors. Self-efficacy has been shown to influence performance on a number of tasks and has been shown to have an effect above and beyond general ability (Bandura, 1990).

For example, Collins (as reported in Bandura, 1990) found that belief in one's mathematical ability influenced performance on a difficult problem-solving task at all levels of ability. She selected children with high, medium, or low levels of mathematical ability and then determined their level of mathematical self-efficacy. Actual ability was an

important factor in performance on the problem-solving task, but at each level of actual ability there was a difference between children who expressed high and low self-efficacy. (I have estimated the following percentages from Bandura's Figure 14.4.) Children with low mathematical ability who expressed high self-efficacy solved 42% of the problems, while those who expressed low self-efficacy solved 19%. There was a 23 unit difference between the two groups. At the medium level of mathematical ability, children who expressed high self-efficacy solved 48% of the problems, and the children who expressed low self-efficacy solved 29%: a difference of 19 units. Among the children with a high level of mathematical ability, there was still a difference between those who expressed high self-efficacy (67% solved) and low self-efficacy (58% solved), but the difference between them was only 9 units.

The results of Collins (as reported in Bandura, 1990) indicate that at all levels of ability, self-efficacy had an effect on the performance of a difficult task. In the present research, it is predicted that high self-efficacy should lead to more judgmental accuracy.

Bandura and Wood (1989) found that participants who were informed that an experimental task was easy and who were given low performance standards showed increasing levels of self-efficacy during the course of their experiment. Conversely, participants who were informed that the task was

difficult and who were given high performance standards showed decreasing levels of self-efficacy over the course of their experiment. In the present research, the judgment tasks are relatively difficult, but no information will be given to participants with respect to how difficult they are. One dimension of difficulty, however, might be whether participants receive feedback concerning their judgments. If participants have no knowledge of how well they are doing, this might make the task more difficult. In the present research, the self-efficacy scales will be administered several times to find whether self-efficacy changes as a result of receiving feedback about one's judgmental accuracy.

# EXPERIMENT 1 METHODS

In Experiment 1, my principle focus was to investigate the effects of receiving feedback versus not receiving feedback on judgments of positive contingencies. Participants completed three sets of seven problems in a discrete-trial format and one set of 14 problems in a summary table format. During the discrete-trial format, half of the participants received feedback after each problem while the other half did not. No feedback was given during the summary table task.

This experiment was also designed to investigate whether increased exposure to a judgment of contingency task would improve judgments. It may be that exposure to a judgment of contingency task and multiple problems will improve judgmental accuracy. In addition, this experiment was designed to investigate whether feedback has an effect on self-efficacy. Last, the inclusion of the summary table task allows for an assessment of whether accuracy on one judgment of contingency task transfers to accuracy on a different contingency task.

## Research Participants

Eighty-six undergraduate students enrolled in introductory psychology courses participated in this experiment to fulfill a requirement. They were randomly assigned to experimental conditions with the restriction that the feedback and no feedback conditions contain an equal

number of participants. Each student was recruited for two hours of participation.

## Materials

The materials consisted of self-efficacy scales, contingency problems, and an open-ended question about judgment strategy.

The self-efficacy scales were adapted from Bandura and Wood (1989). They consisted of four items on which participants reported how confident they were that they could make more accurate judgments than 20%, 40%, 60%, and 80% of the participants in the experiment. Participant responses were based on a nine-point scale (1 = no confidence, 5 = some confidence, 9 = complete confidence). The self-efficacy scales were presented after participants had completed a practice problem and at the conclusion of each set of contingency problems.

The discrete-trial contingency problems were presented by means of a Hypercard program on Macintosh computers. Each contingency problem consisted of 24 three-second trials, with a half-second blank screen between trials. Participants could respond (press the space bar) at any time during the three second trial. At the end of each trial, the screen would either flash or not flash based on the participant's response and the programmed probabilities. At the end of each problem, participants provided a judgment of contingency in response to the question, "What was the effect of your behavior

(pressing and not pressing on the space bar) on the screen's flashing?" Judgments were based on a 201 point scale (-100 = prevents flash from occurring, 0 = has no effect, 100 = causes flash to occur). The Hypercard program recorded each response, outcome, and judgment.

After each judgment of a discrete trial problem (except a practice problem), participants in the feedback condition received information concerning their accuracy. A window appeared which informed them of the actual contingency of the problem and how much their judgment deviated from that value.

The summary table contingency problems were based on materials used by Wasserman and Shaklee (1984). Wasserman and Shaklee presented a problem in which a person (Kim) is trying to find the cause of her radio's occasional buzzing by pressing on one of its internal wires. The table summarizes the number of times that each response-outcome possibility occurred during a given problem. Participants' judgments of these problems were made in response to the question "If you were Kim, what would you conclude was the effect of your behavior (tapping and not tapping on the wire) on the radio's buzzing?", and were based on a 9 point scale (-4 = prevents sound from occurring, 0 = has no effect, 4 = causes sound to occur).

The open-ended question solicited people's judgment strategy for the previous set of problems: "Please describe below how you made your judgments on the last seven problems.

That is, on what did you base your evaluations?" Participants were given a full sheet of paper on which to write their response.

Contingency Problems

Each contingency problem consisted of 24 trials. During each trial there was an opportunity for a response (R or no R) and an outcome (O or no O). There were seven contingency problems in the discrete-trial task. The problems had programmed ΔP values evenly spaced between .00 to 1.00. The problems were also programmed so that there would be an outcome frequency of .50 given a response frequency of .50. The programmed problems are presented in Table 1.

Table 1
Programmed Contingency Problems For The Discrete-Trial Task In Experiment 1

| ΔP | p(O/R) | p(O/no R) |
|---|---|---|
| .00 | .50 | .50 |
| .16 | .58 | .42 |
| .34 | .67 | .33 |
| .50 | .75 | .25 |
| .66 | .83 | .17 |
| .84 | .92 | .08 |
| 1.00 | 1.00 | .00 |

The actual ΔP of the problems could differ from the programmed ΔP values because of the number of trials per problem and the probabilistic nature of the computer program. Whether a flash occurs is based on the programmed probabilities for that problem. If the programmed probability

for an outcome given a response is .67 [p(O/R)=.67], it means that on any given trial when there is a response, there is a .67 chance that the screen will flash. This means that there is a .33 chance that the screen will not flash even though there was a response. With a large number of trials, the actual frequency of an outcome would be very close to the programmed probability, but with a small sample of 24 trials, there is a difference on some problems. In addition, if a participant responds a great deal or very little on a particular problem, the actual ΔP may vary from the programmed ΔP because of the small sample size for p(O/R) or P(O/no R).

The seven problems were presented in five different random orders to assess whether there would be any order or context effects. Analyses of judgmental accuracy as a result of problem order revealed no systematic bias.

In the summary table task, the number of times that each response-outcome possibility occurred during a given problem was summarized for participants. The 14 problems used in this experiment are the 14 problems from Wasserman and Shaklee (1984) that had ΔP values from .00 to 1.00. The problems had outcome frequencies ranging from .125 to 1.00 and had a response frequency of .50. The 14 problems were presented in a single random order to all participants. The problems are listed in Table 2.

Table 2
Contingency Problems Used In The Summary Table Task in Experiment 1

| ΔP | p(O/R) | p(O/no R) | p(O) |
|---|---|---|---|
| 1.00 | 1.00 | .00 | .500 |
| .75 | .75 | .00 | .375 |
| .50 | .50 | .00 | .250 |
| .25 | .25 | .00 | .125 |
| .75 | 1.00 | .25 | .625 |
| .50 | .75 | .25 | .500 |
| .25 | .50 | .25 | .375 |
| .00 | .25 | .25 | .250 |
| .50 | 1.00 | .50 | .750 |
| .25 | .75 | .50 | .625 |
| .00 | .50 | .50 | .500 |
| .25 | 1.00 | .75 | .875 |
| .00 | .75 | .75 | .750 |
| .00 | 1.00 | 1.00 | 1.000 |

Note. Table is based on Wasserman and Shaklee's (1984) Table 1.

## Procedure

Participants were randomly assigned to a feedback or a no feedback condition. They were given the following instructions for the discrete-trial task at the beginning of the experiment (adapted from Wasserman et al., [1983] and Wasserman & Shaklee [1984]).

> The aim of this experiment is to see how people judge the relationship between their actions and the consequences of those actions. In the seven problems that follow, the same basic question is posed: What is the relation between your pressing the space bar and the occurrence of a brief flashing on the computer screen? The seven problems differ only in the particular relationship between your pressing and the occurrence of the flash. For each of the seven problems, please rate the degree to which your pressing affects the rate of the screen's flashing, from "prevents the flash from occurring" to "causes the flash to occur."

> Each of the seven problems will take about 2 minutes. Each problem consists of 24 three-second trials. During each three-second trial you can either press the space bar or you can refrain from pressing it. At the end of each trial the screen may flash or it may not. Depending on your response and the screen's outcome, there are four response-outcome possibilities: Press-Flash, Press-No Flash, No Press-Flash, and No Press-No Flash. Each trial is separated by a half-second of blank computer screen.
>
> To make an accurate judgment you will need to notice what happens when you press the space bar and what happens when you don't press it. It will be to your advantage to press the space bar on about half of the 24 trials.

After participants read these instructions, they completed a practice problem, made their judgment of the problem, and were given an opportunity to ask the experimenter questions about the task. After their questions had been addressed, participants completed the pre-task self-efficacy scale, judged the seven contingency problems, completed the post-task self-efficacy scale, and answered the open-ended question about judgment strategy. After a short break, participants judged the second set of seven problems, completed another post-task self-efficacy scale, and answered the open-ended question about judgment strategy. Participants went through these steps again in conjunction with the third set of problems. After participants finished their third short break, they were given the following instructions for the summary table judgment task (adapted from Wasserman & Shaklee, 1984).

> The aim of this experiment is to see how people judge the relationship between their actions and the consequences of those actions. In the 14 problems that follow, the same basic question is posed: What is the relation between Kim's tapping on the wire of a malfunctioning radio and the occurrence of a brief buzzing sound that the radio occasionally emits. The 14 problems differ only in the particular relationship between Kim's tapping and the occurrence of the sound. For each of the 14 problems, please rate the degree to which Kim's tapping affects the rate of the radio's buzzing, from "prevents the sound from occurring" to "causes the sound to occur." It is more important to work through the problems carefully and methodically than to give quick and offhand reactions.

The next page presented the 14 problems in summary table format. The response-outcome possibilities were listed along with the number of times that each occurred for a given problem. At the top of the page the following instructions appeared.

> After buying a new radio, Kim finds that it emits a brief buzzing sound every so often. Kim finds this buzzing sound annoying and decides to find its cause. Removing the back of the radio, Kim suspects that a wire may be loose. Kim chooses a wire and taps on it a number of times in order to see if this has any effect on the buzzing sound. In the table below, Kim's tapping on the wire and the radio's buzzing have been summarized into four response-outcome possibilities. The number of times that each response-outcome possibility occurred for each problem is listed below.

At this point participants judged the 14 problems, completed the post-task self-efficacy scale, and completed the open-ended question about judgment strategy. Participants were then debriefed about the aims of this experiment.

# RESULTS AND DISCUSSION OF EXPERIMENT 1

## Data Screening

The data of any participant who did not complete the experiment in an appropriate manner were excluded from the analyses. This rule excluded the data of participants who did not complete all of the materials, who completed the materials out of sequence, or who did not respond appropriately on the discrete-trial contingency problems (either responding on every trial or not responding at all).

I collected data for Experiment 1 until there was complete data from 86 people. The judgmental accuracy of each participant was assessed as follows. The absolute difference between a participant's judgment of a problem and the problem's actual contingency was calculated for every problem. These absolute difference scores were then averaged for each of the problem sets. This produced four mean difference scores for each participant (mean difference score on problem set 1, problem set 2, etc.).

Data screening revealed some rather extreme difference scores. One participant in the feedback condition performed poorly on the first two problem sets, with mean difference scores of 56.86 and 55.43, respectively. On the third problem set, her performance deteriorated, with a mean difference score of 103.57. Because this individual was in the feedback condition, she probably knew that she was performing worse in

the third problem set than in the previous problem sets. I am at a loss to explain this, and can only comment that it is no small feat to be off by an average of 103.57 points on a 201 point scale.

Because of this individual's extreme scores, I implemented an across-the-board selection criteria. I excluded the data of any participant whose set 1, set 2, or set 3 mean difference score was more than three standard deviations from the grand mean. (Because the mean difference scores for the Kim problem were based on a different scale, they were not used in this aspect of the data screening.) Six participants were excluded from further analyses. The 80 remaining participants are evenly divided between the feedback and no feedback conditions (40 each), with 16 men and 24 women in each condition.

## Analysis of the Discrete-Trial Task Mean Difference Scores on Problem Sets 1, 2, and 3

A repeated-measures analysis of variance (ANOVA) was performed using SPSS (1990). The between-subject variables were condition (feedback and no feedback) and participant's sex (male and female). The repeated-measures variable was mean absolute difference scores for the three discrete-trial problem sets. The means and standard deviations are presented in Table 3. The significance level for all statistical tests is $p < .05$. Any test with $p > .10$ will not be interpreted.

Table 3
Mean Absolute Difference Scores for the Discrete-Trial Task by Sex (Experiment 1)

| | Condition | | | |
|---|---|---|---|---|
| | Feedback (n = 40) | | No Feedback (n = 40) | |
| Problem Set | M | SD | M | SD |
| 1 | | | | |
| Male [1] | 19.04 | 9.17 | 23.37 | 10.98 |
| Female [2] | 20.27 | 9.49 | 20.50 | 7.44 |
| 2 | | | | |
| Male | 16.01 | 11.09 | 22.80 | 9.31 |
| Female | 17.46 | 7.23 | 17.51 | 7.01 |
| 3 | | | | |
| Male | 12.18 | 8.31 | 22.63 | 10.29 |
| Female | 14.48 | 6.75 | 20.59 | 10.11 |

[1] There were 16 men per condition.
[2] There were 24 women per condition.

The ANOVA revealed that there was no reliable main effect or interaction involving the sex independent variable. The sex by condition interaction was the only statistical test to produce a $p$ value $< .10$, $F(1, 76) = 2.85$, $p < .10$. All further analyses collapsed across the sex variable.

A second repeated-measures ANOVA was performed that examined experimental condition (between-subject variable) and mean absolute difference scores (repeated-measures variable). The means and standard deviations are presented in Table 4. Participants in the feedback condition were more accurate in their judgments than participants in the no feedback condition, $F(1, 78) = 7.87$, $p < .01$. There was an improvement in judgmental accuracy over the three problem

sets, $F$ (2, 156) = 4.37, $p < .05$. The mean difference scores of the no feedback condition remained around 20 for all three problem sets, but the mean difference scores of the feedback condition improve. This difference in improvement between participants in the two conditions is confirmed as a significant interaction between condition and problem set, $F$ (2, 156) = 4.00, $p < .05$.

Table 4
Mean Absolute Difference Scores for the Discrete-Trial Task (Experiment 1)

| | Condition | | | |
|---|---|---|---|---|
| | Feedback (n = 40) | | No Feedback (n = 40) | |
| Problem Set | M | SD | M | SD |
| 1 | 19.78 | 9.27 | 21.65 | 9.00 |
| 2 | 16.88 | 8.87 | 19.63 | 8.32 |
| 3 | 13.56 | 7.40 | 21.40 | 10.10 |

One of the principal questions behind this experiment was whether feedback would improve judgmental accuracy. Research by Jenkins and Ward (1965) suggested that feedback might not improve judgmental accuracy, while research by Clark and Benassi (in press) suggested that it might. The results indicate that the feedback provided in Experiment 1 improved judgmental accuracy as evidenced by the decrease in mean difference scores.

A second question of this experiment was whether practice at judging a number of contingency problems would

improve judgmental accuracy. Participants became better judges of contingency over the three problem sets. This effect was due to the considerable improvement shown by participants in the feedback condition. The mean difference scores of participants in the no feedback condition showed no systematic improvement. This pattern of results suggests that practice can improve judgmental accuracy when combined with feedback, but that practice alone does not improve judgmental accuracy.

Mean Judgments of Contingency by Level of Contingency

As discussed in the Introduction, the psychophysical function for the judgment of contingency suggests that there is a difference in discriminability between different levels of contingency. Past research has shown that participants tend to underestimate the objective degree of contingency, producing a shallow function. This underestimation of the degree of contingency is the most pronounced for contingencies between -.50 and .50.

To assess whether this pattern of judgment held for the present study, I examined mean judgments of contingency by the level of contingency in the following manner. I categorized the problems into seven groups according to actual $\Delta P$. The midpoint of each group's interval was the value of one of the programmed contingencies (.00, .17, .33, .50, .67, .83, 1.00). The lowest group's interval included contingencies from -.08 to .08. This interval includes some

negative contingencies, but none that are far removed from noncontingency. Problems with an actual contingency below -.08 were not the focus of Experiment 1 and did not occur with great frequency. There were only 115 problems below this level out of 1680 problems in Experiment 1 (6.85%). These problems are not included in the present analyses.

The highest group's interval included contingencies from .92 to 1.00. This interval is only half that of the other intervals, but this group still had a large number of problems. There were 339 problems in this group out of 1680 problems in Experiment 1 (20.18%). (The midpoint of this group is .96, but I refer to it as 1.00 group because [a] it is the programmed contingency and [b] most of the problems in this group were 1.00 contingencies.)

I calculated the mean actual contingency and the mean judgment of contingency for each of the groups. These data are presented in Tables 5, 6, and 7, for the three problem sets. These data are also presented graphically in Figures 2, 3, and 4. No statistical tests have been performed because of nonindependence of the data points. Each participant contributed up to seven judgments per problem set. These data are presented only for descriptive purposes. (Note that the number of data points per group and per condition vary. This is because the groups were defined by actual $\Delta P$ and any problems with $\Delta P$ lower than -.08 were excluded.)

Table 5
Mean Actual Contingency and Mean Judgment of Contingency for Problem Set 1 by Level of Contingency (Experiment 1)

| Condition | | | | | |
|---|---|---|---|---|---|
| Feedback (n = 260) | | | No Feedback (n = 258) | | |
| n | Mean ΔP | Mean Judgment | n | Mean ΔP | Mean Judgment |
| 36 | .02 | 17.94 | 28 | .00 | 9.82 |
| 32 | .18 | 31.13 | 34 | .18 | 6.56 |
| 30 | .34 | 29.90 | 34 | .33 | 21.32 |
| 40 | .50 | 51.63 | 39 | .50 | 39.15 |
| 30 | .64 | 49.67 | 37 | .69 | 76.16 |
| 37 | .82 | 81.51 | 26 | .82 | 76.73 |
| 55 | .98 | 90.89 | 60 | .98 | 96.48 |

Figure 2
Mean Actual Contingency and Mean Judgment of Contingency for Problem Set 1 (Experiment 1)

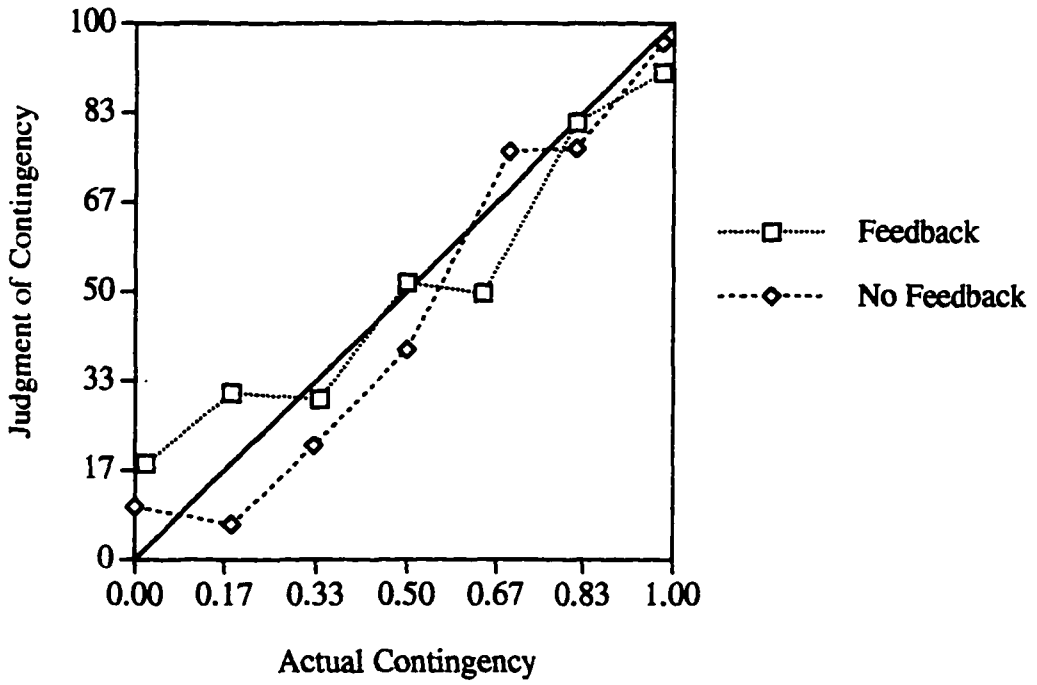


Table 6
Mean Actual Contingency and Mean Judgment of Contingency for Problem Set 2 by Level of Contingency (Experiment 1)

| Condition | | | | | |
|---|---|---|---|---|---|
| Feedback (n = 258) | | | No Feedback (n = 264) | | |
| n | Mean ΔP | Mean Judgment | n | Mean ΔP | Mean Judgment |
| 22 | -.01 | 6.68 | 29 | .01 | 6.69 |
| 33 | .19 | 22.91 | 39 | .18 | 9.26 |
| 35 | .34 | 40.97 | 26 | .34 | 19.12 |
| 44 | .52 | 44.02 | 44 | .50 | 43.96 |
| 36 | .67 | 63.11 | 44 | .67 | 58.96 |
| 25 | .84 | 73.56 | 27 | .83 | 77.48 |
| 63 | .98 | 94.67 | 55 | .99 | 95.76 |

Figure 3
Mean Actual Contingency and Mean Judgment of Contingency for Problem Set 2 (Experiment 1)

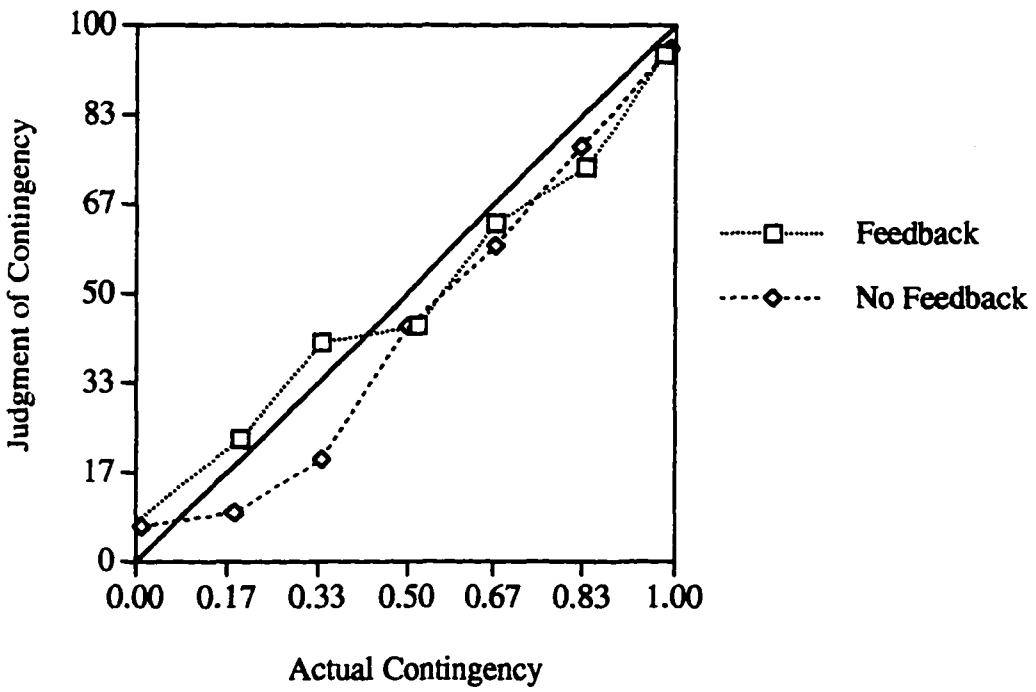
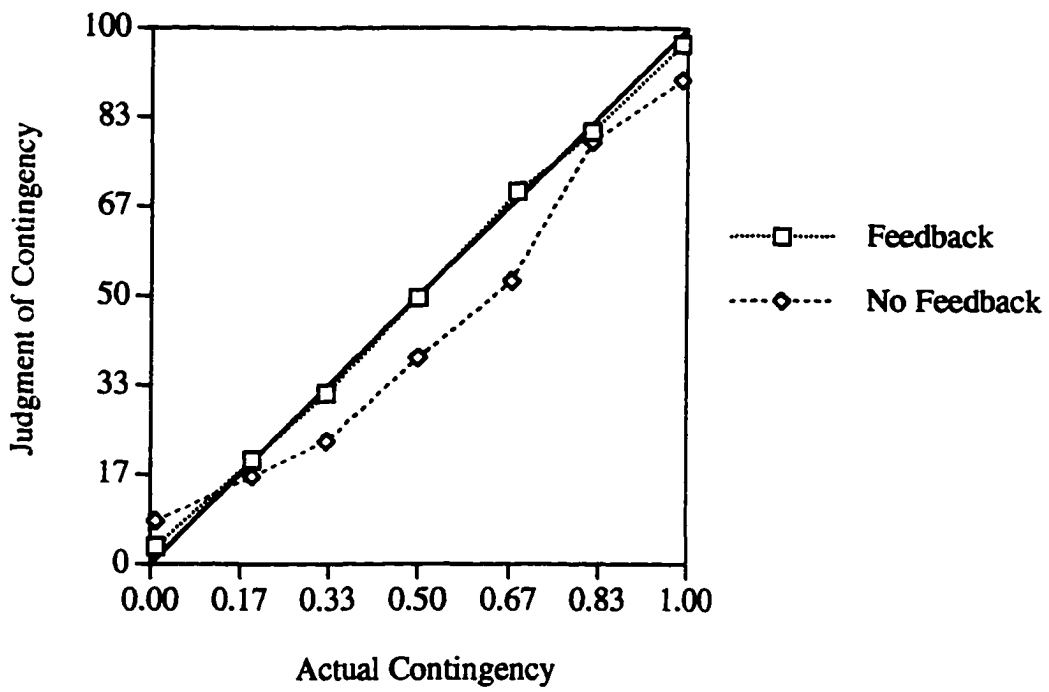


Table 7
Mean Actual Contingency and Mean Judgment of Contingency for Problem Set 3 by Level of Contingency (Experiment 1)

| Condition | | | | | |
|---|---|---|---|---|---|
| Feedback (n = 258) | | | No Feedback (n = 267) | | |
| n | Mean ΔP | Mean Judgment | n | Mean ΔP | Mean Judgment |
| 22 | .01 | 3.27 | 32 | .01 | 7.94 |
| 36 | .19 | 19.31 | 37 | .19 | 16.24 |
| 31 | .33 | 31.87 | 27 | .33 | 22.89 |
| 52 | .50 | 49.77 | 49 | .50 | 38.51 |
| 32 | .68 | 69.47 | 36 | .67 | 52.89 |
| 31 | .82 | 80.68 | 34 | .82 | 78.82 |
| 54 | .99 | 97.11 | 52 | .99 | 90.35 |

Figure 4
Mean Actual Contingency and Mean Judgment of Contingency for Problem Set 3 (Experiment 1)



The mean judgment of contingency data show that participants' judgments of contingency more closely approximate the actual contingencies over the course of the three problem sets. This is especially true for the data of participants in the feedback condition.

In addition, these data indicate that participants' mean judgments of contingency are more accurate for contingencies above .50 than they are for contingencies below .50. This finding confirms what was described in the Introduction as the typical psychophysical function for the judgment of contingency.

The difference between the judgments of higher versus lower levels of contingency can be illustrated by calculating the slope and intercept for the least squares regression line for all of the contingencies, the contingencies between .00 and .50, and the contingencies between .50 and 1.00. The results are presented in Table 8. A pattern of veridical judgments would have a slope of 100 and an intercept of 0.

Table 8
Slope and Intercept for the Least Squares Regression Line Between Mean Actual Contingency and Mean Judgment of Contingency (Experiment 1)

| | Coefficients for Contingencies Between | | |
|---|---|---|---|
| | .00 and 1.00 | .00 and .50 | .50 and 1.00 |
| | Problem Set 1 | | |
| Feedback | | | |
| Slope | 75.94 | 61.70 | 93.95 |
| Intercept | 12.59 | 16.64 | -.84 |
| No Feedback | | | |
| Slope | 99.74 | 62.24 | 111.30 |
| Intercept | -3.22 | 3.52 | -10.92 |
| | Problem Set 2 | | |
| Feedback | | | |
| Slope | 84.22 | 75.11 | 103.69 |
| Intercept | 6.96 | 9.07 | -9.06 |
| No Feedback | | | |
| Slope | 96.64 | 74.76 | 106.65 |
| Intercept | -4.13 | .66 | -10.85 |
| | Problem Set 3 | | |
| Feedback | | | |
| Slope | 97.37 | 94.44 | 96.08 |
| Intercept | 1.12 | 1.54 | 2.31 |
| No Feedback | | | |
| Slope | 88.42 | 61.10 | 111.42 |
| Intercept | -.24 | 5.82 | -17.71 |

In interpreting these data, consider first the data of participants in the feedback condition. On problem set 1, the slope and intercept are quite different for the low versus high ranges of contingency. The slope for the low range of contingencies is shallow (61.70) and the intercept is high (16.64). The slope for the high range of contingencies is steeper (93.95) than for the low range of contingencies and the intercept is near 0 (-.84). The slopes have a difference of 32.25 and the intercepts have a difference of 17.48. On

problem set 3, there is only a small difference between the slopes and intercepts for the low and high contingencies. There is a difference of 1.64 for the slopes and .77 for the intercepts. In addition to the two regression lines being similar, note that the slopes are relatively close to 100 and that the intercepts are both close to 0.

Next consider the data of the participants in no feedback condition. On problem set 1, there is a considerable difference between the slope and intercept for the low and high ranges of contingency. The slope for the low range of contingencies is shallow (62.24) and the intercept is low (3.52). The slope for the high range of contingencies is steeper than 100 (111.30) and the intercept is below 0 (-10.92). The slopes have a difference of 49.06 and the intercepts have a difference of 14.44. On problem set 3, the pattern of judgment is basically the same. There is a shallow slope for the low range of contingencies (61.10) with a low intercept (5.82). There is a slope greater than 100 for the high range of contingencies (111.42) with an intercept well below 0 (-17.71). Unlike the feedback condition, there is still a large difference between the slopes and intercepts for the low and the high ranges of contingency.

It appears that the effect of feedback was to bring participants' judgments of contingency closer in line with veridical judgment. This effect produced similar slopes and intercepts for the low and high ranges of contingency.

Without feedback, there was no systematic coming together of judgments for the low versus high ranges of contingency. This is another source of evidence indicating that judgmental accuracy does not improve with practice alone.

Difference Score by Level of Contingency

The first analysis I performed (repeated-measures ANOVA) examined judgmental accuracy as indicated by mean absolute difference scores. These mean absolute difference scores represented the average difference between judged contingency and actual contingency for the seven problems of each problem set. This measure provided a measure of a participant's average accuracy across all levels of contingency.

The second analysis I performed examined mean judgments of contingency by level of contingency. The mean judgments of contingency provided a measure of judgmental accuracy for each level of actual contingency. These mean judgments are scores that collapse across participants.

In the present set of analyses, I again rely on mean absolute difference scores between judged contingency and actual contingency, but these mean difference scores collapse across individuals and represent mean absolute difference scores for each level of contingency. These mean absolute difference scores represent the average difference between participants' judgments of contingency and the actual contingencies for each level of contingency. These analyses

allow a second way to address the differential discriminability of contingencies.

Using the same contingency groups as in the analysis of mean judgments of contingency by actual contingency, I calculated the mean absolute difference score for each contingency group. These data are presented in Tables 9, 10, and 11 for problem sets 1, 2, and 3, respectively. The data are presented graphically in Figures 5, 6, and 7. No statistical tests have been performed because of nonindependence of the data points. Each participant contributed up to seven judgments per problem set. These data are presented only for descriptive purposes. (Note that the number of data points per group and per condition vary. This is because the groups were defined by actual $\Delta P$ and any problems with $\Delta P$ lower than -.08 were excluded.)

Table 9
Mean Absolute Difference Scores by Level of Contingency for Problem Set 1 (Experiment 1)

| Contingency Group | Condition | | | | | |
|---|---|---|---|---|---|---|
| | Feedback (n = 260) | | | No Feedback (n = 258) | | |
| | n | M | SD | n | M | SD |
| .00 | 36 | 22.61 | 24.00 | 28 | 29.43 | 26.61 |
| .17 | 32 | 23.16 | 16.68 | 34 | 24.82 | 24.06 |
| .33 | 30 | 24.57 | 17.13 | 34 | 30.18 | 20.09 |
| .50 | 40 | 21.08 | 13.56 | 39 | 30.05 | 23.14 |
| .67 | 30 | 29.23 | 24.30 | 37 | 16.87 | 9.45 |
| .83 | 37 | 13.73 | 15.99 | 26 | 16.23 | 18.46 |
| 1.00 | 55 | 8.40 | 21.59 | 60 | 3.32 | 7.40 |

Figure 5
Mean Absolute Difference Score by Level of Contingency for Problem Set 1 (Experiment 1)

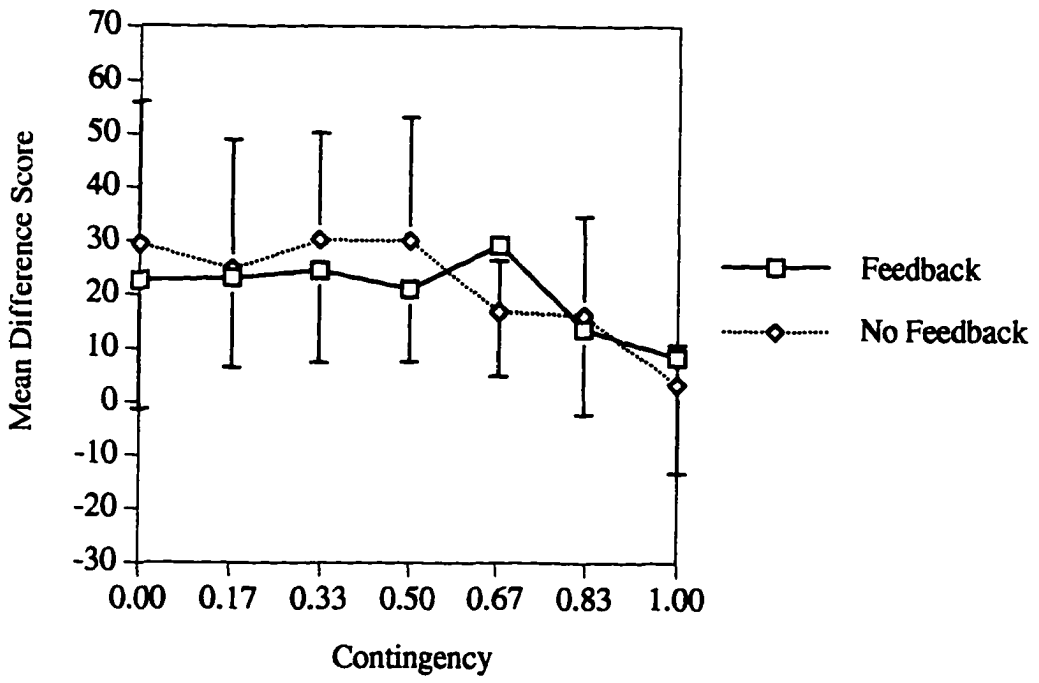


Table 10
Mean Absolute Difference Scores by Level of Contingency for Problem Set 2 (Experiment 1)

| Contingency Group | Condition | | | | | |
|---|---|---|---|---|---|---|
| | Feedback (n = 258) | | | No Feedback (n = 264) | | |
| | n | M | SD | n | M | SD |
| .00 | 22 | 19.41 | 20.40 | 29 | 18.38 | 21.56 |
| .17 | 33 | 15.70 | 15.83 | 39 | 26.67 | 24.72 |
| .33 | 35 | 25.91 | 15.72 | 26 | 29.04 | 11.40 |
| .50 | 44 | 20.11 | 14.81 | 44 | 22.80 | 21.60 |
| .67 | 36 | 20.17 | 16.35 | 44 | 23.75 | 21.98 |
| .83 | 25 | 16.24 | 21.66 | 27 | 15.70 | 20.13 |
| 1.00 | 63 | 4.78 | 13.27 | 55 | 4.11 | 14.55 |

Figure 6
Mean Absolute Difference Score by Level of Contingency for Problem Set 2 (Experiment 1)

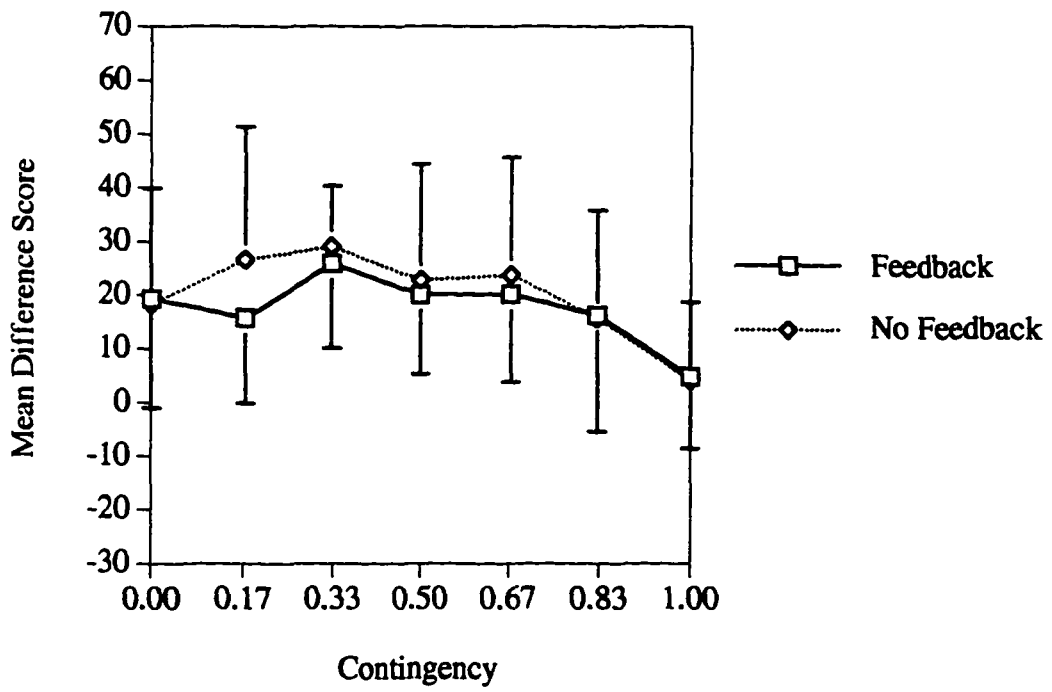
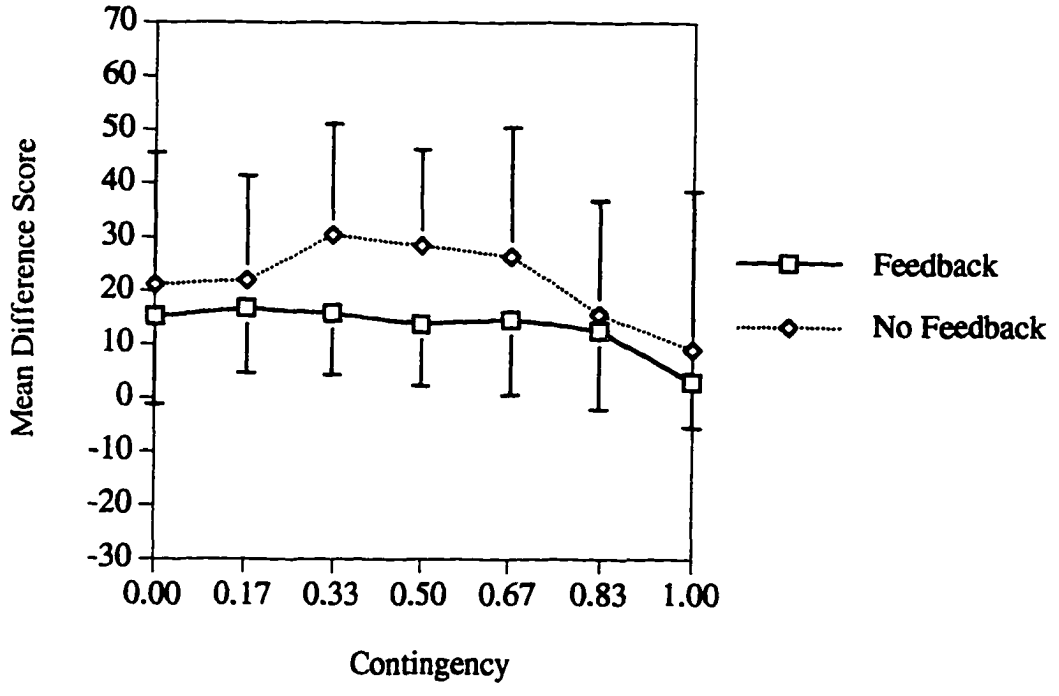


Table 11
Mean Absolute Difference Scores by Level of Contingency for Problem Set 3 (Experiment 1)

| Contingency Group | Condition | | | | | |
|---|---|---|---|---|---|---|
| | Feedback (n = 258) | | | No Feedback (n = 267) | | |
| | n | M | SD | n | M | SD |
| .00 | 22 | 15.27 | 16.53 | 32 | 21.19 | 24.54 |
| .17 | 36 | 16.78 | 12.08 | 37 | 22.03 | 19.37 |
| .33 | 31 | 15.84 | 11.54 | 27 | 30.44 | 20.70 |
| .50 | 52 | 13.87 | 11.56 | 49 | 28.49 | 17.77 |
| .67 | 32 | 14.72 | 14.16 | 36 | 26.47 | 24.03 |
| .83 | 31 | 12.68 | 14.85 | 34 | 15.65 | 21.18 |
| 1.00 | 54 | 2.98 | 8.50 | 52 | 9.04 | 29.62 |

Figure 7
Mean Absolute Difference Score by Level of Contingency for Problem Set 3 (Experiment 1)



These data indicate that the smallest mean difference scores occur for contingencies in the .83 and 1.00 groups. This finding suggests that those problems with an objective contingency greater than .75 support the most judgmental accuracy (.75 is the lower cutoff for the .83 group). Participants are not as accurate in their judgments of problems that are less contingent than $\Delta P = .75$. This confirms that there is a difference in the discriminability of contingencies across levels of contingency.

One pattern in the data was unexpected. In Figure 7, the mean difference scores for participants in the no feedback condition are lower for contingencies in the .00 and .17 groups than they are for the middle three groups of contingencies (.33, .50, and .67). There is no similar improvement in the judgments of participants in the feedback condition. This difference led me to believe that participants in the no feedback condition judged "0" as a default for any problem that was closer to being noncontingent than being perfectly contingent. If this is the case, it would explain the relatively small mean difference scores for the .00 and .17 contingency groups and the relatively large difference scores for the .33, .50, and .67 contingency groups. Participants in the feedback condition would not be expected to show this same pattern of judging "0"s because they knew that relatively few of the problems were actual "0"s.

To find out whether my speculation was correct, I found the number of "0" judgments at each level of contingency. The results for problem set 3 are presented in Table 12.

Table 12
Number of "0" Judgments and Total Judgments for Each Level of Contingency in Problem Set 3 (Experiment 1)

| Contingency Group | Condition | | | | | |
|---|---|---|---|---|---|---|
| | Feedback | | | No Feedback | | |
| | "0"s | Total | % of Total | "0"s | Total | % of Total |
| .00 | 4 | 22 | 18 | 12 | 32 | 38 |
| .17 | 3 | 36 | 8 | 13 | 37 | 35 |
| .33 | 3 | 31 | 10 | 9 | 27 | 33 |
| .50 | 1 | 52 | 2 | 7 | 49 | 14 |
| .67 | 0 | 32 | 0 | 4 | 36 | 11 |
| .83 | 0 | 31 | 0 | 0 | 34 | 0 |
| 1.00 | 0 | 54 | 0 | 1 | 52 | 2 |

In the feedback condition, only 18% of the judgments in the .00 contingency group were "0"s. This is a marked contrast to the no feedback condition in which 38% of the judgments for this level of contingency were "0"s. In addition, participants in the no feedback condition judged that over one-third of the problems in the .17 and .33 groups were "0"s. These participants even judged that over 10% of the problems in the .50 and .67 contingency groups were "0"s.

The use of "0" as a default judgment would help account for the no feedback condition's having smaller mean absolute difference scores for the contingencies in the .00 and .17

groups and having relatively larger mean absolute difference scores for contingencies in the .33, .50, and .67 groups.

Response Rate

Wasserman et al. (1983) found a difference in judgmental accuracy as a function of participants' response rate. In their Experiment 1, a median split of participants according to their mean response rate produced four groups: press-low, tap-low, press-high, and tap-high. The four groups had mean response probabilities of .17, .23, .37, and .44, respectively. (Mean response probability refers to the probability of a response on a given trial.) The press-high and tap-high groups provided judgments of contingency that were consistent with the actual contingencies for the problems. The press-low and tap-low groups were less accurate in their judgments of contingency.

To assess whether this pattern of judgment is supported by the present research, I performed a three-way split of participants according to their mean response rate. The top third of participants had a mean response rate above 13.74 (mean response probability = .63). The bottom third of participants had a mean response rate below 12.66 (mean response probability = .49.). I then compared the mean absolute difference score on the three problem sets for the high and low response thirds. These means and standard deviations are presented in Table 13.

Table 13
Mean Absolute Difference Scores on the Discrete-Trial Task for Participants with a High or Low Response Rate (Experiment 1)

| | Condition | | | |
|---|---|---|---|---|
| | Feedback (n = 29) [1] | | No Feedback (n = 23) [2] | |
| Problem Set | M | SD | M | SD |
| 1 | | | | |
| High Response Rate | 17.40 | 6.73 | 22.77 | 9.12 |
| Low Response Rate | 19.27 | 8.96 | 20.96 | 8.58 |
| 2 | | | | |
| High Response Rate | 17.64 | 8.50 | 20.07 | 7.80 |
| Low Response Rate | 12.57 | 7.46 | 17.70 | 6.08 |
| 3 | | | | |
| High Response Rate | 14.89 | 7.50 | 23.42 | 10.27 |
| Low Response Rate | 11.46 | 6.98 | 19.35 | 8.49 |

[1] In the feedback condition, there were 14 participants with a high response rate and 15 participants with a low response rate.
[2] In the no feedback condition, there were 12 participants with a high response rate and 11 participants with a low response rate.

A repeated-measures ANOVA was performed that examined condition (feedback and no feedback) and response rate (high and low) as between-subject variables. The repeated-measures variable was mean absolute difference scores on the three problem sets. There was no systematic difference in judgmental accuracy associated with response rate, $F$ (1, 48) = 2.26, $p < .10$). None of the interactions involving response rate were significant (all $p$ values > .10).

The present research did not produce the same pattern of results that Wasserman et al. (1983) found. I suspect that the reason for this difference stems from the different response patterns in our experiments. In Wasserman et al.'s

Experiment 1, the press-low and the tap-low groups had very low mean response probabilities (.17 and .23, respectively). The press-high and tap-high groups had mean response probabilities that were within 15 units of an optimal rate of response (.37 and .44, respectively). In my Experiment 1, both the low and the high response rate groups had a mean response probability that was within 15 units of an optimal rate of response (.49 and .63, respectively).

In sum, the difference between what Wasserman et al. (1983) found for response rate and what I found may be due to the fact that most of the participants in my Experiment 1 responded at a near optimal rate. In the Wasserman et al.'s Experiment 1 this was not the case.

Self-Efficacy

One of the questions of this research was whether receiving feedback about one's performance would influence self-efficacy. A mean self-efficacy score was obtained for each participant on each of the self-efficacy scales. Because this analysis examines whether feedback influences self-efficacy, the data from the first and fifth self-efficacy scales are not included. There was no feedback for either the practice problem (which preceded the first self-efficacy scale) or the summary table task (which preceded the fifth self-efficacy scale).

The mean self-efficacy scores for the second, third, and fourth self-efficacy scales are presented in Table 14. Higher scores indicate higher levels of self-efficacy.

Table 14
Mean Self-Efficacy Scores for Self-Efficacy Scales 2, 3, and 4 by Sex (Experiment 1)

| | Condition | | | |
|---|---|---|---|---|
| | Feedback (n = 40) | | No Feedback (n = 40) | |
| Self-Efficacy Scale | M | SD | M | SD |
| 2 | | | | |
| Male [1] | 5.72 | 1.43 | 5.61 | 1.14 |
| Female [2] | 5.30 | 1.37 | 5.24 | 1.28 |
| 3 | | | | |
| Male | 6.36 | 1.73 | 5.64 | 1.60 |
| Female | 5.43 | 1.70 | 5.33 | 1.78 |
| 4 | | | | |
| Male | 6.52 | 1.57 | 5.27 | 1.60 |
| Female | 5.52 | 1.48 | 5.00 | 1.59 |

[1] There were 16 men per condition.
[2] There were 24 women per condition.

A repeated-measures ANOVA was performed with condition (feedback and no feedback) and sex (male and female) as the between-subject variables. Mean self-efficacy score was the repeated-measures variable. Male participants reported higher levels of self-efficacy than female participants, but this test failed to reach the $p < .05$ level, $F(1, 76) = 3.41$, $p < .07$. None of the interactions involving the sex variable had $p$ values $< .10$.

A second repeated-measures ANOVA was performed which collapsed across participant's sex. The mean self-efficacy scores and standard deviations are presented in Table 15.

Table 15
Mean Self-Efficacy Scores for Self-Efficacy Scales 2, 3, and 4 (Experiment 1)

| | Condition | | | |
|---|---|---|---|---|
| | Feedback (n = 40) | | No Feedback (n = 40) | |
| Self-Efficacy Scale | M | SD | M | SD |
| 2 | 5.47 | 1.39 | 5.39 | 1.23 |
| 3 | 5.80 | 1.75 | 5.46 | 1.35 |
| 4 | 5.92 | 1.58 | 5.11 | 1.58 |

There was no difference in self-efficacy as a result of receiving feedback, $F$ (1, 78) = 1.97, $p > .10$. There was no main effect for the repeated-measures variable, $F$ (2, 156) = 1.11, $p > .10$. There was, however, a significant interaction between condition and the repeated-measures variable, $F$ (2, 156) = 3.79, $p < .05$. For the feedback condition, mean self-efficacy increased over the course of the experiment. For the no feedback condition, mean self-efficacy decreased.

These analyses reveal two interesting patterns. First, there was a tendency for male participants to report higher levels of self-efficacy than female participants. This is interesting in light of the fact that there was no difference in judgmental accuracy as a function of participants' sex. Second, there was no main effect for condition (feedback or

no feedback) on self-efficacy. The interaction between condition and the repeated-measures variable revealed that participants in the feedback reported more self-efficacy as the experiment progressed.

The above analyses address whether receiving feedback influences self-efficacy. I conducted another analysis to assess whether participants with high self-efficacy were better judges of contingency than participants with low self-efficacy. I calculated mean self-efficacy scores for each participant over the second, third, and fourth self-efficacy scales. Participants were selected for this analysis by their having a mean self-efficacy score on these scales that was in the top third or bottom third of all participants. The top third of participants had a mean self-efficacy score above 6.20. (The mean self-efficacy score for these 27 participants was 6.97. [*SD* = .54].) The bottom third of participants had a mean self-efficacy score below 4.95. (The mean self-efficacy score for these 27 participants was 4.10 [*SD* = .77].) I calculated the mean absolute difference scores on the discrete-trial task for these participants (see Table 16).

Table 16
Mean Absolute Difference Scores on the Discrete-Trial Task for Participants with High or Low Self-Efficacy (Experiment 1)

| | Condition | | | |
|---|---|---|---|---|
| | Feedback (n = 30) [1] | | No Feedback (n = 24) [2] | |
| Problem Set | M | SD | M | SD |
| 1 | | | | |
| High Self-Efficacy | 20.04 | 9.82 | 24.30 | 14.42 |
| Low Self-Efficacy | 17.92 | 10.23 | 20.20 | 7.41 |
| 2 | | | | |
| High Self-Efficacy | 13.74 | 9.75 | 22.44 | 10.67 |
| Low Self-Efficacy | 19.61 | 7.14 | 19.21 | 7.86 |
| 3 | | | | |
| High Self-Efficacy | 10.69 | 7.76 | 25.22 | 12.52 |
| Low Self-Efficacy | 16.55 | 6.43 | 19.96 | 8.10 |

[1] In the feedback condition, there were 18 participants with high self-efficacy and 12 participants with low self-efficacy.
[2] In the no feedback condition, there were 9 participants with high self-efficacy and 15 participants with low self-efficacy.

A repeated-measures ANOVA was performed with condition (feedback and no feedback) and self-efficacy (high and low) as the between-subject variables. Mean absolute difference scores on the three problem sets of the discrete-trial task was the repeated-measures variable. There was no difference in judgmental accuracy for participants in the high versus low self-efficacy groups, $F$ (1, 50) = .07, $p$ > .10. The interaction between condition and self-efficacy approached the $p$ < .05 level of significance, $F$ (1, 50) = 3.68, $p$ < .07. Participants with high self-efficacy in the feedback condition showed more judgmental accuracy than participants with low self-efficacy. In the no feedback condition it was

just the opposite, participants with high self-efficacy showed less judgmental accuracy than participants with low self-efficacy. None of the other interactions was significant (all $p$ values > .10). There was no main effect for the repeated-measures variable of judgmental accuracy over the three problem sets ($p$ > .10).

Bandura (1986) and Collins (as reported in Bandura, 1990) report that self-efficacy has been shown to influence performance above and beyond the influence of general ability. Collins found that children with high self-efficacy solved more math problems than children with low self-efficacy. In the present research, participants with high self-efficacy were better judges of contingency, but only in the feedback condition. In the no feedback condition, participants with high self-efficacy were worse judges of contingency than participants with low self-efficacy. Why were these participants poor judges of contingency? One reason might be that these participants were confident in their ability and their judgments to the extent that they were not attentive to the differences between the problems they performed.

## Analysis of the Summary Table Task

### Mean Difference Scores on the Summary Table Task

An initial assessment of judgmental accuracy on the summary table task revealed that there was no difference between the two experimental conditions. Both conditions had

the same mean absolute difference score ($M$ = 1.03) and similar standard deviations (feedback $SD$ = .60, no feedback $SD$ = .51; $n$ = 40 for both conditions). (This is a small mean difference score, but it comes from judgments made on a nine point scale. As a ratio of mean difference score to the size of the judgment scale, this is equivalent to a mean difference score of 23 on the 201 point scale used in the discrete-trial task.) It appears that feedback did not have a systematic effect on judgmental accuracy on a new task.

Correlation Between Mean Absolute Difference Scores on Problem Set 3 of the Discrete-Trial Task and the Summary Table Task

There was no between-condition difference in judgmental accuracy, but it may be the case that participants who were accurate on the discrete-trial task were also accurate on the summary table task. To assess whether there was any consistency between performance on the two tasks at the level of the individual, mean absolute difference scores on problem set 3 (the last of the discrete-trial problem sets) and problem set 4 (the summary table task) were correlated. The correlation approached the .05 level, $r$ = .21, $p < .07$, $n$ = 80. This suggests that participants who were relatively accurate on one of the tasks were also relatively accurate on the other task.

Taken separately, neither the feedback ($r$ = .21, $p$ > .10, $n$ = 40) nor the no feedback conditions ($r$ = .25, $p$ > .10, $n$ = 40) approached significance due to the sample size.

Comparison to Wasserman and Shaklee (1984)

The above analyses show little evidence of a transfer of judgmental accuracy from one contingency task to another. If there were a transfer, the mean absolute difference scores on the summary task would be expected to be smaller for participants in the feedback condition. The discrete-trial task may still have had an effect on participants' judgments on the summary task. Whether this is the case can be found by comparing participant's judgments of these problems to the judgments obtained by Wasserman and Shaklee (1984). Table 17 presents mean judgments and standard deviations of the summary table problems from the present research (collapsed across the feedback and no feedback conditions).

Table 17
Mean and Standard Deviation (in Parentheses) of Judgments to the Summary Table Contingency Problems (Experiment 1)

| | Contingency | | | | | |
|---|---|---|---|---|---|---|
| p(O) [1] | .00 | .25 | .50 | .75 | 1.00 | Row Average |
| .125 | | -1.06 (1.82) | | | | -1.06 |
| .250 | -.58 (1.44) | | .28 (2.11) | | | -.15 |
| .375 | | .08 (1.41) | | 1.29 (2.12) | | .69 |
| .500 | .01 (.41) | | 1.80 (1.22) | | 3.33 (1.72) | 1.71 |
| .625 | | 1.16 (1.10) | | 2.55 (1.45) | | 1.86 |
| .750 | .35 (.81) | | 2.39 (1.11) | | | 1.37 |
| .875 | | 1.65 (1.35) | | | | 1.65 |
| 1.000 | .26 (1.02) | | | | | .26 |
| Column Average | .01 | .46 | 1.49 | 1.92 | 3.33 | |

Note. n = 80
[1] Denotes the probability of an outcome.

Table 18 presents Wasserman and Shaklee's (1984) data from their Experiment 4 for the same contingency problems.

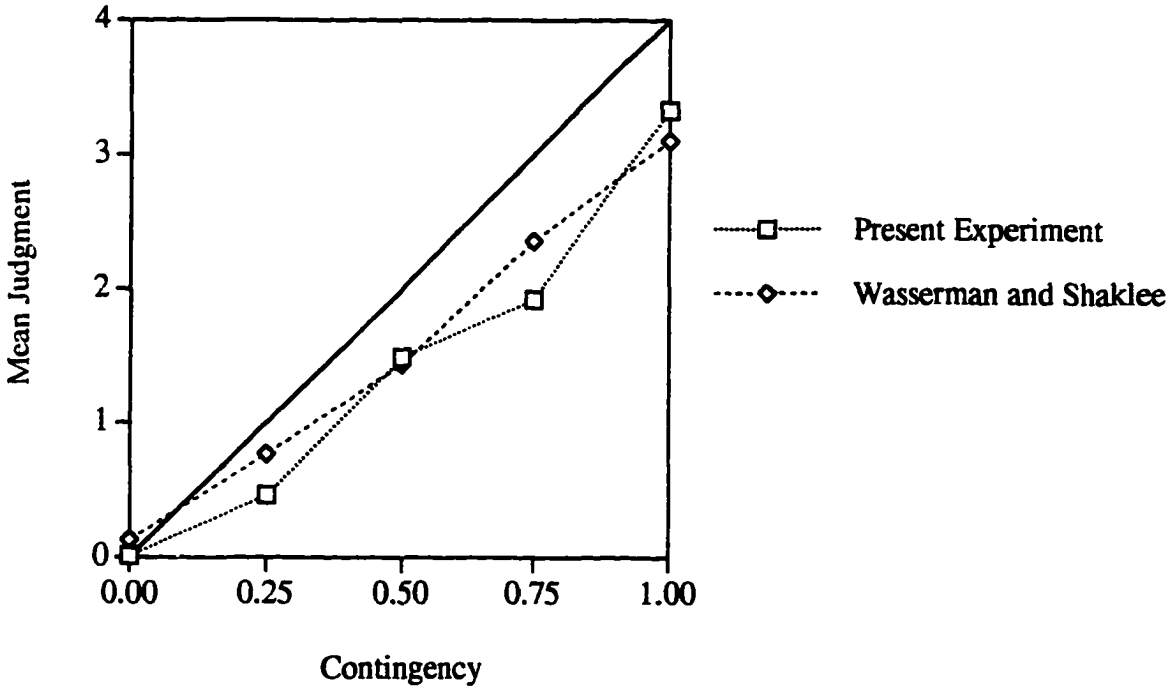
Table 18
Mean and Standard Deviation (in Parentheses) of Judgments to the Summary Table Contingency Problems from Wasserman and Shaklee (1984, Experiment 4)

| | Contingency | | | | | |
|---|---|---|---|---|---|---|
| p(O) [1] | .00 | .25 | .50 | .75 | 1.00 | Row Average |
| .125 | | -.25 (1.82) | | | | -.25 |
| .250 | -.38 (1.35) | | .68 (1.79) | | | .15 |
| .375 | | .53 (1.40) | | 1.93 (1.99) | | 1.23 |
| .500 | -.03 (.47) | | 1.65 (1.35) | | 3.10 (1.77) | 1.57 |
| .625 | | 1.20 (1.03) | | 2.78 (.88) | | 1.99 |
| .750 | .58 (.92) | | 1.98 (1.35) | | | 1.28 |
| .875 | | 1.58 (1.20) | | | | 1.58 |
| 1.000 | .35 (1.26) | | | | | .35 |
| Column Average | .13 | .77 | 1.44 | 2.36 | 3.10 | |

Note. n = 40
[1] Denotes the probability of an outcome.

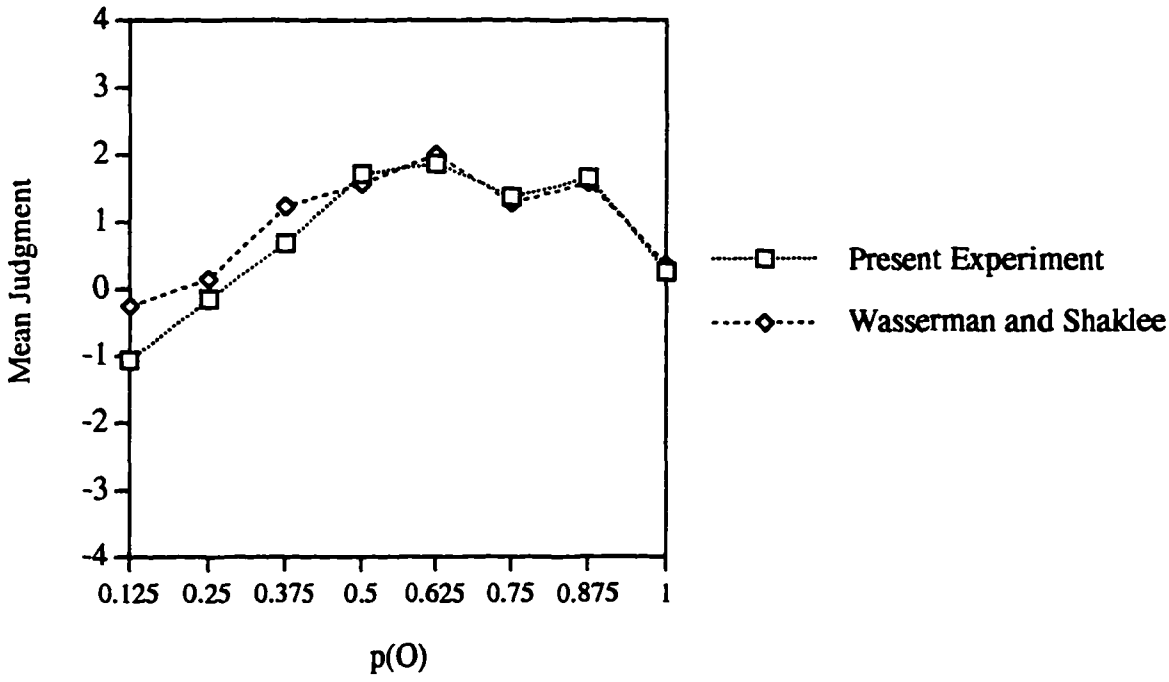
These tables reveal a similar pattern of judgment. First, they show that participants in both experiments accurately scaled contingencies. In other words, participants judged low contingencies to be lower in value than higher contingencies. Figure 8 presents the mean contingency judgment for each level of contingency (column averages) for Tables 17 and 18.

Figure 8
Mean Judgments of Contingency for Each Level of Contingency in the Summary Table Task from the Present Experiment and from Wasserman and Shaklee (1984, Experiment 4)




Second, participants in both experiments showed a tendency to give higher judgments when there was a higher outcome frequency. Figure 9 presents the mean contingency judgment for each level of outcome frequency (row averages) for Tables 17 and 18.

Figure 9
Mean Judgments of Contingency for Each Level of Outcome Frequency in the Summary Table Task from the Present Experiment and from Wasserman and Shaklee (1984, Experiment 4 )




The similarity between the data of the present experiment and those of Wasserman and Shaklee (1984) is quite clear. It appears that having been through the discrete-trial task did not have much effect on participants' judgments of the summary table problems.

# EXPERIMENT 2 METHODS

The purpose of Experiment 2 was to assess whether I could replicate the effects of feedback and practice found in Experiment 1 when using a more difficult set of contingency problems. In Experiment 1 I employed contingency problems with $\Delta P$ values between .00 and 1.00. In Experiment 2 I employed problems with $\Delta P$ values between -1.00 and .00.

I did not collect data on self-efficacy, judgment strategies, or the transfer of judgmental accuracy from one task to another. By not including these elements from Experiment 1, this experiment required less time from each participant. In addition, the length of each trial was reduced from three seconds to two seconds and participants were not given breaks between the three problem sets. With these modifications to Experiment 2, I could recruit participants for one hour of participation. This allowed me to collect more judgment of contingency data with my allotment from the Psychology 401 Participant Pool.

## Research Participants

Eighty-six undergraduates enrolled in introductory psychology participated in this experiment to fulfill a course requirement. They were randomly assigned to experimental conditions with the restriction that the feedback and no feedback conditions contain an equal number

of participants. Each student was recruited for one hour of participation.

## Materials

The materials consisted only of instructions for the discrete-trial contingency task. The instructions were modified slightly from those of Experiment 1 to reflect the changes in Experiment 2.

The contingency problems were presented by means of a Hypercard program on Macintosh computers. Each contingency problem consisted of 24 two-second trials, with a half-second blank screen between trials. Participants could respond (press the space bar) at any time during the two-second trial. At the end of each trial, the screen would either flash or not flash based on a participant's response and the programmed probabilities. At the end of each problem, participants provided a judgment of contingency in response to the question, "What was the effect of your behavior (pressing and not pressing on the space bar) on the screen's flashing?" Judgments were based on a 201 point scale (-100 = prevents flash from occurring, 0 = has no effect, 100 = causes flash to occur). The Hypercard program recorded each response, outcome, and judgment.

After each judgment of a discrete trial problem (except a practice problem), participants in the feedback condition received information concerning their accuracy. A window

appeared which informed them of the actual contingency of the problem and how much their judgment deviated from that value.

The seven contingency problems in Experiment 2 had programmed ΔP values evenly spaced between -1.00 and .00. The problems were also programmed so that there would be an outcome frequency of .50 given a response frequency of .50. The programmed problems are presented in Table 19.

Table 19
Programmed Contingency Problems For The Discrete-Trial Task In Experiment 2

| ΔP | p(O/R) | p(O/no R) |
|---|---|---|
| -1.00 | .00 | 1.00 |
| -.84 | .08 | .92 |
| -.66 | .17 | .83 |
| -.50 | .25 | .75 |
| -.34 | .33 | .67 |
| -.16 | .42 | .58 |
| .00 | .50 | .50 |

The seven problems were presented in five different random orders to assess whether there would be any order or context effects. Analyses of judgmental accuracy as a result of problem order revealed no systematic bias.

# RESULTS AND DISCUSSION OF EXPERIMENT 2

## Data Screening

As in Experiment 1, the data of any participant who did not complete the experiment in an appropriate manner were excluded from the analyses. I collected data for Experiment 2 until there was complete data from 86 people. Judgmental accuracy was assessed as follows. The absolute difference between a participant's judgment of a problem and the problem's actual contingency was calculated for every problem. These absolute difference scores were then averaged for each of the three problem sets.

I applied the same across-the-board selection criterion as in Experiment 1. That is, I excluded the data of any participant whose set 1, set 2, or set 3 mean absolute difference score was more than three standard deviations from the grand mean. Data from six participants were excluded from further analyses. The 80 remaining participants were evenly divided between the feedback and no feedback conditions (40 each), with 16 men and 24 women in each condition.

## Analysis of the Discrete-Trial Task

### Mean Absolute Difference Scores on Problem Sets 1, 2, and 3

A repeated-measures analysis of variance (ANOVA) was performed using SPSS (1990). The between-subject variables were condition (feedback and no feedback) and participant's sex (male and female). The repeated-measures variable was

mean absolute difference scores on the three problem sets. The means and standard deviations are presented in Table 20. As was the case with Experiment 1, the significance level for all statistical tests is $p < .05$. Any test with $p > .10$ will not be interpreted.

Table 20
Mean Absolute Difference Scores for the Discrete-Trial Task by Sex (Experiment 2)

| | Condition | | | |
|---|---|---|---|---|
| | Feedback (n = 40) | | No Feedback (n = 40) | |
| Problem Set | M | SD | M | SD |
| 1 | | | | |
| Male [1] | 20.46 | 11.71 | 20.53 | 10.80 |
| Female [2] | 21.79 | 10.21 | 25.09 | 9.06 |
| 2 | | | | |
| Male | 18.34 | 9.02 | 19.38 | 9.73 |
| Female | 15.74 | 5.93 | 26.36 | 10.12 |
| 3 | | | | |
| Male | 18.76 | 6.71 | 20.87 | 10.33 |
| Female | 13.82 | 5.04 | 25.87 | 9.29 |

[1] There were 16 men per condition.
[2] There were 24 women per condition.

The ANOVA revealed that participants in the feedback condition were more accurate in their judgments than participants in the no feedback condition, $F(1, 76) = 10.02$, $p < .01$. There was no main effect for sex ($F(1, 76) = 1.25$, $p > .10$), but there was a significant interaction between sex and condition, $F(1, 76) = 6.09$, $p < .05$. The women in the feedback condition were the most accurate group of participants; the women in the no feedback condition were the

least accurate group of participants. The men in both conditions performed similarly.

There was no main effect for the repeated-measures variable, $F$ (2, 152) = 2.00, $p > .10$. This indicates that there was no across-the-board improvement in judgmental accuracy over the three problem sets. The interaction between the repeated-measures variable and condition approached significance, $F$ (2, 152) = 2.78, $p < .07$. Participants in the feedback condition showed a tendency to improve in judgmental accuracy over the three problem sets. There was no significant interaction between the repeated-measures variable and sex, $F$ (2, 152) = .79, $p > .10$.

For purposes of comparison with Experiment 1, Table 21 presents mean absolute difference scores for Experiment 2 collapsed across participant's sex. This table presents information comparable to Table 4.

Table 21
Mean Absolute Difference Scores for the Discrete-Trial Task (Experiment 2)

| | Condition | | | |
|---|---|---|---|---|
| | Feedback (n = 40) | | No Feedback (n = 40) | |
| Problem Set | M | SD | M | SD |
| 1 | 21.26 | 10.71 | 23.26 | 9.92 |
| 2 | 16.78 | 7.33 | 23.57 | 10.43 |
| 3 | 15.79 | 6.19 | 23.87 | 9.90 |

The data in Table 21 reveal a pattern similar to that of Table 4. In the feedback condition there is an improvement in participants' mean absolute difference score over the three problem sets. There is no comparable improvement in the judgments of participants in the no feedback condition. A comparison of these tables also reveals that mean absolute difference scores are smaller in Experiment 1 than in Experiment 2. This finding suggests that the positive contingency problems used in Experiment 1 supported more judgmental accuracy than the negative contingency problems used in Experiment 2.

One of the principal questions behind Experiment 2 was whether I would find the same pattern of results as in Experiment 1. Just as in Experiment 1, feedback improved judgmental accuracy in Experiment 2. Further, practice alone did not improve judgmental accuracy for the no feedback condition.

Mean Judgments of Contingency by Level of Contingency

As discussed in the Introduction, the psychophysical function for the judgment of contingency indicates that there is a difference in discriminability between different levels of contingency. Past research has shown that participants tend to underestimate the objective degree of contingency, producing a shallow function. This underestimation of the degree of contingency is the most pronounced for contingencies between -.50 and .50. This pattern of judgment

was found in Experiment 1. To assess whether it would also be found in Experiment 2, I again examined mean judgments of contingency by the level of contingency.

I categorized the problems into seven groups by actual contingency. The midpoint of each group was the value of one of the programmed contingencies (-1.00, -.83, -.67, -.50, -.33, -.17, .00). The .00 group's interval included contingencies between -.08 and .08. This interval includes some positive contingencies, but none that are far removed from noncontingency. Problems with an actual contingency above .08 were not the focus of this experiment and did not occur with great frequency. There were only 108 problems with an actual contingency above .08 out of 1680 problems in Experiment 2 (6.43%). These problems are not included in the present analyses.

The -1.00 group's interval included contingencies between -1.00 and -.92. This interval is only half that of the other intervals, but this group still had a large number of problems. There were 323 problems in this group out of 1680 problems in Experiment 2 (19.23%). (The midpoint of this interval is -.96, but I refer to it as the -1.00 group because [a] it is the programmed contingency and [b] most of the problems in this group were -1.00 contingencies.)

I calculated the mean actual contingency and the mean judgment of contingency for each of the groups. These data are presented in Tables 22, 23, and 24 for the three problem

sets. These data are also presented graphically in Figures 10, 11, and 12. For ease of comparison to Figures 2, 3, and 4, the axes of Figures 10, 11, and 12 have been reversed so that an underestimation of contingency falls below the diagonal that represents veridical judgment. No statistical tests have been performed because of nonindependence of the data points. Each participant contributed up to seven judgments per problem set. (Note that the number of data points per group and per condition vary. This is because the groups were defined by actual $\Delta P$ and any problems with $\Delta P$ greater than .08 were excluded.)

Table 22
Mean Actual Contingency and Mean Judgment of Contingency for Problem Set 1 by Level of Contingency (Experiment 2)

| Condition | | | | | |
|---|---|---|---|---|---|
| Feedback (n = 261) | | | No Feedback (n = 261) | | |
| n | Mean ΔP | Mean Judgment | n | Mean ΔP | Mean Judgment |
| 49 | -.99 | -95.18 | 52 | -.99 | -99.00 |
| 50 | -.82 | -71.48 | 31 | -.84 | -77.13 |
| 34 | -.68 | -61.18 | 36 | -.67 | -55.11 |
| 37 | -.51 | -38.76 | 44 | -.50 | -28.52 |
| 32 | -.34 | -27.75 | 35 | -.34 | -3.34 |
| 32 | -.19 | -5.34 | 37 | -.19 | -8.95 |
| 27 | .01 | 4.44 | 26 | -.02 | -5.39 |

Figure 10
Mean Actual Contingency and Mean Judgment of Contingency for Problem Set 1 (Experiment 2)

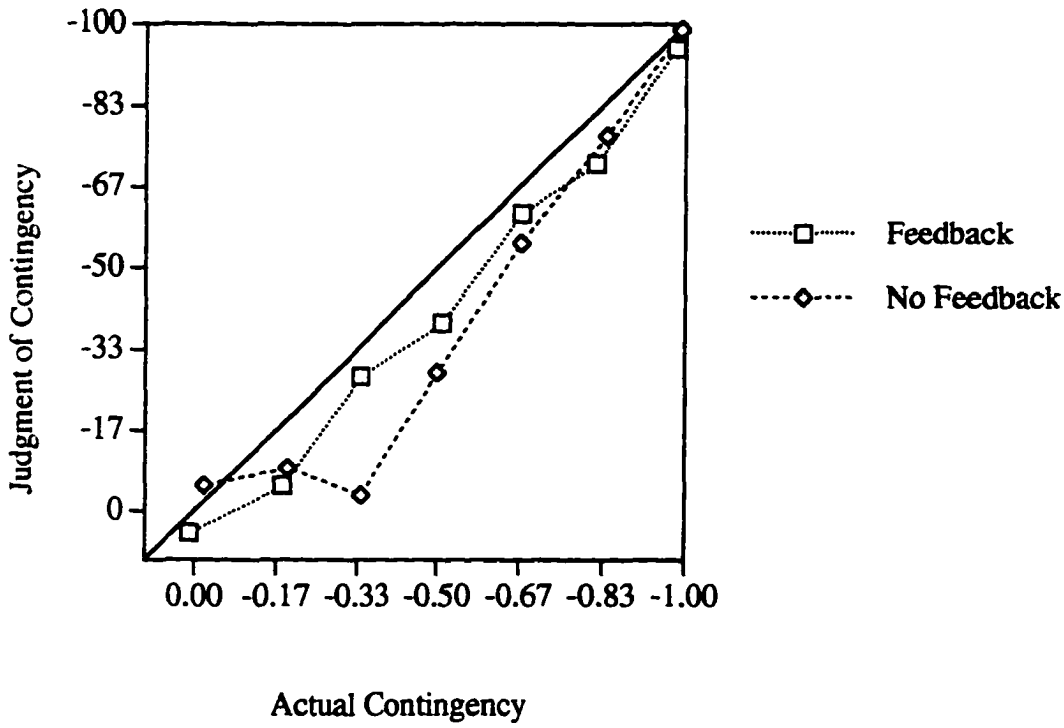


Table 23
Mean Actual Contingency and Mean Judgment of Contingency for Problem Set 2 by Level of Contingency (Experiment 2)

| Condition | | | | | |
|---|---|---|---|---|---|
| Feedback (n = 268) | | | No Feedback (n = 260) | | |
| n | Mean ΔP | Mean Judgment | n | Mean ΔP | Mean Judgment |
| 56 | -.99 | -95.04 | 51 | -.99 | -99.10 |
| 34 | -.82 | -77.29 | 33 | -.82 | -66.88 |
| 38 | -.66 | -58.92 | 38 | -.68 | -42.55 |
| 37 | -.50 | -42.08 | 43 | -.51 | -27.19 |
| 31 | -.35 | -30.29 | 29 | -.34 | -20.86 |
| 43 | -.19 | -18.26 | 40 | -.17 | -5.23 |
| 29 | .00 | 5.07 | 26 | .00 | 1.08 |

Figure 11
Mean Actual Contingency and Mean Judgment of Contingency for Problem Set 2 (Experiment 2)

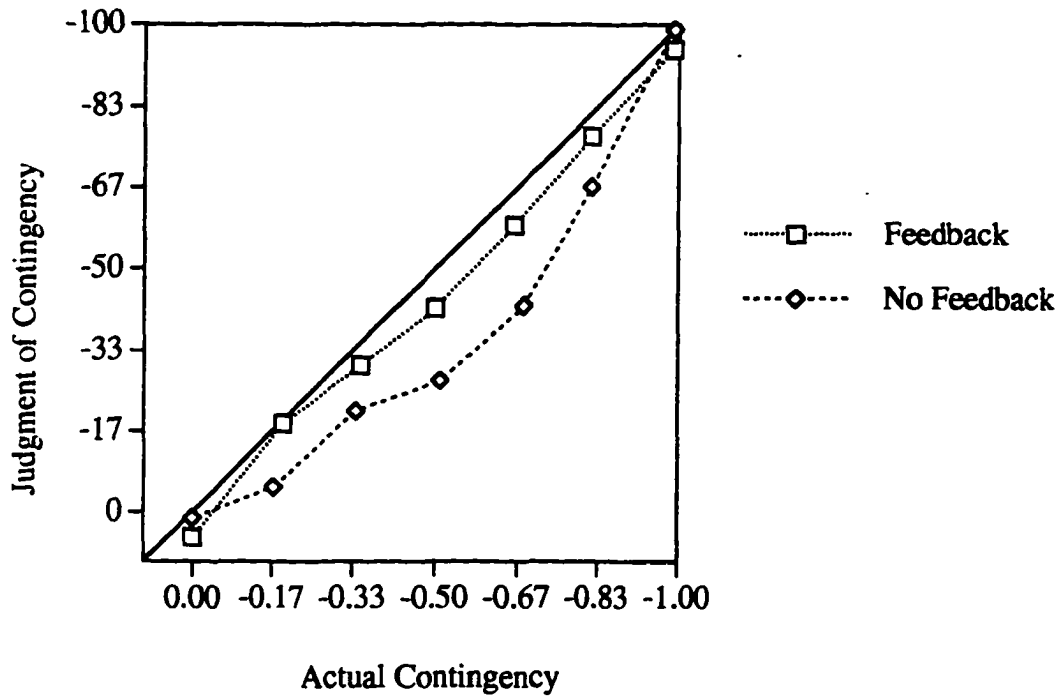
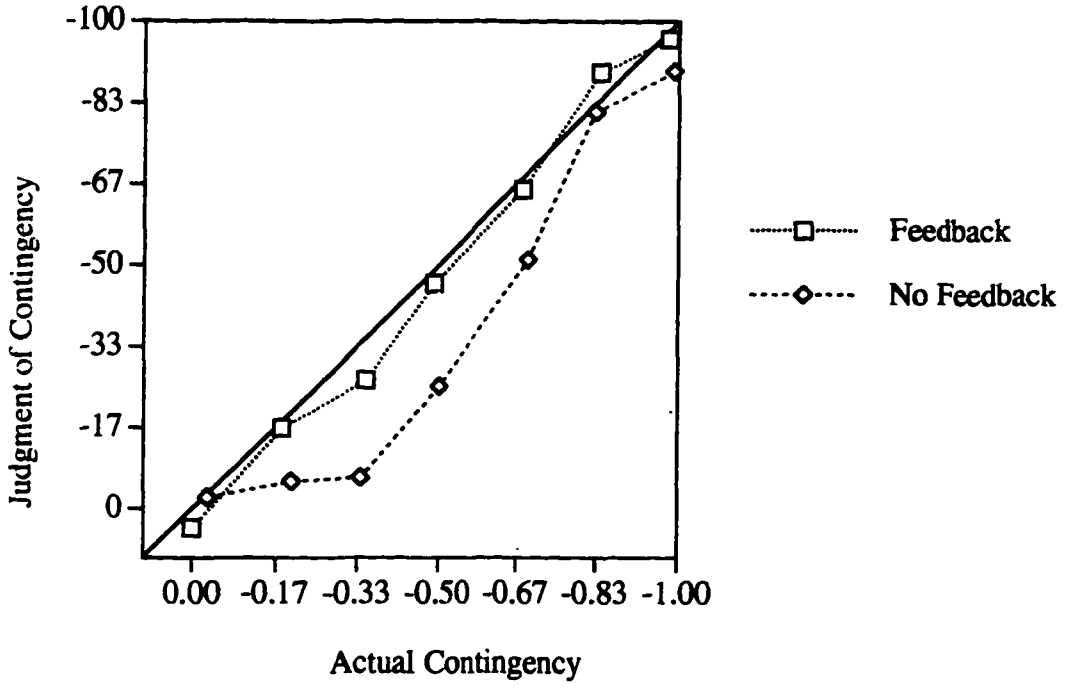


Table 24
Mean Actual Contingency and Mean Judgment of Contingency for Problem Set 3 by Level of Contingency (Experiment 2)

| Condition | | | | | |
|---|---|---|---|---|---|
| Feedback (n = 258) | | | No Feedback (n = 264) | | |
| n | Mean ΔP | Mean Judgment | n | Mean ΔP | Mean Judgment |
| 57 | -.98 | -96.49 | 58 | -.99 | -89.83 |
| 24 | -.84 | -89.54 | 27 | -.83 | -81.48 |
| 36 | -.68 | -65.58 | 42 | -.69 | -51.19 |
| 47 | -.49 | -46.32 | 37 | -.50 | -25.16 |
| 30 | -.35 | -26.50 | 34 | -.34 | -6.53 |
| 37 | -.18 | -16.70 | 39 | -.20 | -5.69 |
| 27 | .00 | 3.89 | 27 | -.03 | -2.44 |

Figure 12
Mean Actual Contingency and Mean Judgment of Contingency for Problem Set 3 (Experiment 2)



These data show that the mean judgments of participants in the feedback condition more closely approximate the actual contingencies than the judgments of participants in the no feedback condition. The judgments of participants in the feedback condition of Experiment 1 were also closer to being veridical than the judgments of participants in the no feedback condition.

In Experiment 1, participants had a tendency to underestimate the degree of contingency for problems that had relatively little contingency ($\Delta P < .50.$). They had less of a tendency to underestimate problems that had relatively more contingency ($\Delta P > .50.$) In Experiment 2, it appears that participants had a similar tendency to underestimate problems with relatively little contingency. Whether there is a systematic difference in the judgment of lower contingency problems and higher contingency problems can be assessed by calculating the slope and intercept for the least squares regression line for all of the contingencies, the contingencies between -.50 and .00, and the contingencies between -1.00 and -.50. These results are presented in Table 25. A pattern of veridical judgments would have a slope of 100 and an intercept of 0.

Table 25
Slope and Intercept for the Least Squares Regression Line for Mean Actual Contingency and Mean Judgment of Contingency (Experiment 2)

| | Coefficients for Contingencies Between | | |
|---|---|---|---|
| | -1.00 and .00 | -.50 and .00 | -1.00 and -.50 |
| | Problem Set 1 | | |
| Feedback | | | |
| Slope | 100.22 | 86.97 | 113.91 |
| Intercept | 8.11 | 5.39 | 18.79 |
| No Feedback | | | |
| Slope | 103.38 | 39.99 | 142.31 |
| Intercept | 13.07 | -.92 | 42.23 |
| | Problem Set 2 | | |
| Feedback | | | |
| Slope | 98.70 | 93.51 | 109.65 |
| Intercept | 4.24 | 2.92 | 13.15 |
| No Feedback | | | |
| Slope | 95.97 | 59.19 | 150.23 |
| Intercept | 11.12 | 2.16 | 54.11 |
| | Problem Set 3 | | |
| Feedback | | | |
| Slope | 105.49 | 97.16 | 106.63 |
| Intercept | 4.89 | 3.31 | 5.28 |
| No Feedback | | | |
| Slope | 102.19 | 44.39 | 139.84 |
| Intercept | 14.47 | 1.77 | 42.93 |

In interpreting these data, consider first the feedback condition. On problem set 1, the slope and intercept are quite different for the two ranges of contingency. The slope for the low range of contingencies is shallow (86.97) and the intercept is low (5.39). The slope for the high range of contingencies is steep (113.91) and the slope is high (18.79). The slopes have a difference of 26.94 and the intercepts have a difference of 13.40. On problem set 3,

there is a smaller difference between the slopes and intercepts for the low and high contingencies (a difference of 9.47 for the slopes and 1.97 for the intercepts). In addition to the slopes and intercepts being similar, note that the slopes are closer to 100 than they were on problem set 1 and that the intercepts are both close to 0.

In contrast, note that there is no comparable improvement in the no feedback condition. On problem set 1, there is a considerable difference between the slopes and intercepts for the low and high ranges of contingencies. The slope for the low range of contingencies is shallow (39.99) and the intercept is low (-.92). The slope for the high range of contingencies is steep (142.31) and the intercept is high (42.23). The slopes have a difference of 102.32 and the intercepts have a difference of 43.15. On problem set 3, this pattern is unchanged. The slope for the low range of contingencies is shallow (44.39) and the intercept is low (1.77). The slope for the high range of contingencies is steep (139.84) and the intercept is high (42.93). The slopes have a difference of 95.45 and the intercepts have a difference of 41.16.

As was found in Experiment 1, it appears that the effect of feedback was to bring participants' judgments in line with veridical judgment. This effect made the slopes and intercepts similar for the low and high ranges of

contingency. Without feedback, there was no coming together of judgments for the low and high ranges of contingency.

This mean judgment of contingency data support the established function for the judgment of contingency. Overall, people tend to underestimate the objective value of contingencies. This is most pronounced for contingencies between -.50 and .50. This data supports that there is a difference in the differential discriminability of contingencies.

Difference Score by Level of Contingency

In the present set of analyses, I examine mean absolute difference scores that represent the difference between judged contingency and actual contingency. These mean absolute difference scores collapse across individuals and show the mean absolute difference scores for contingencies of different levels. These analyses allow for a second way of assessing the differential discriminability of contingencies.

Using the same contingency groups as in the mean judgment of contingency analyses, I calculated the mean absolute difference score for each contingency group. These data are presented in Tables 26, 27, 28. The data are also presented graphically in Figures 13, 14, and 15. For ease of comparison with figures 5, 6, and 7, the horizontal axis representing contingency has been reversed. Noncontingent problems are represented at the left of the figure and

perfectly contingent problems are represented at the right of the figure.

No statistical tests have been performed because of nonindependence of the data. Each participant contributed up to seven judgments per problem set. These data are presented only for descriptive purposes. (Note that the number of data per group and per condition vary. This is because the groups are defined by actual $\Delta P$ and any problems with $\Delta P$ greater than .08 were excluded.)

Table 26
Mean Absolute Difference Scores by Level of Contingency for Problem Set 1 (Experiment 2)

| Contingency Group | Condition | | | | | |
|---|---|---|---|---|---|---|
| | Feedback (n = 261) | | | No Feedback (n = 261) | | |
| | n | M | SD | n | M | SD |
| -1.00 | 49 | 5.12 | 17.22 | 52 | .92 | 3.74 |
| -.83 | 50 | 22.88 | 39.56 | 31 | 22.65 | 37.24 |
| -.67 | 34 | 24.15 | 26.16 | 36 | 33.69 | 36.01 |
| -.50 | 37 | 29.30 | 26.21 | 44 | 37.71 | 18.08 |
| -.33 | 32 | 26.81 | 18.01 | 35 | 33.20 | 18.24 |
| -.17 | 32 | 26.09 | 18.06 | 37 | 26.46 | 18.17 |
| .00 | 27 | 22.82 | 22.41 | 26 | 10.96 | 19.08 |

Figure 13
Mean Absolute Difference Score by Level of Contingency for Problem Set 1 (Experiment 2)

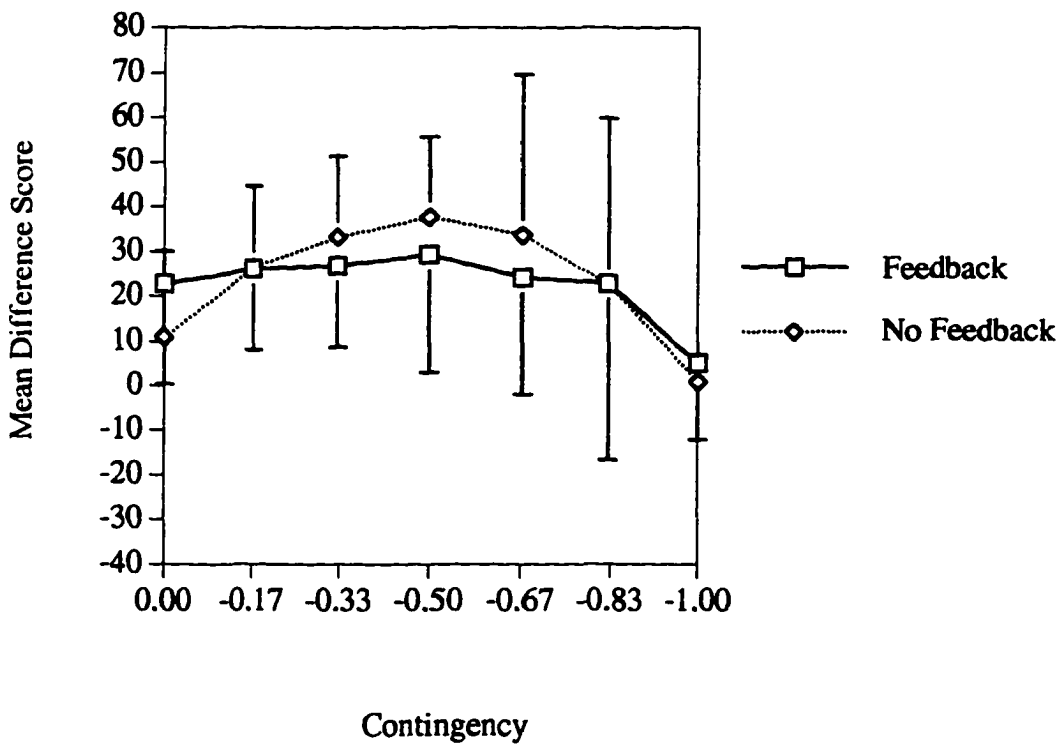


Table 27
Mean Absolute Difference Scores by Level of Contingency for Problem Set 2 (Experiment 2)

| Contingency Group | Condition | | | | | |
|---|---|---|---|---|---|---|
| | Feedback (n = 268) | | | No Feedback (n = 260) | | |
| | n | M | SD | n | M | SD |
| -1.00 | 56 | 4.84 | 26.70 | 51 | .90 | 3.79 |
| -.83 | 34 | 14.94 | 17.94 | 33 | 27.70 | 41.43 |
| -.67 | 38 | 17.24 | 19.67 | 38 | 38.95 | 28.21 |
| -.50 | 37 | 22.27 | 15.31 | 43 | 32.61 | 20.37 |
| -.33 | 31 | 23.52 | 17.76 | 29 | 34.38 | 15.28 |
| -.17 | 43 | 20.98 | 12.67 | 40 | 24.40 | 18.74 |
| .00 | 29 | 17.14 | 15.25 | 26 | 13.85 | 18.36 |

Figure 14
Mean Absolute Difference Score by Level of Contingency for Problem Set 2 (Experiment 2)

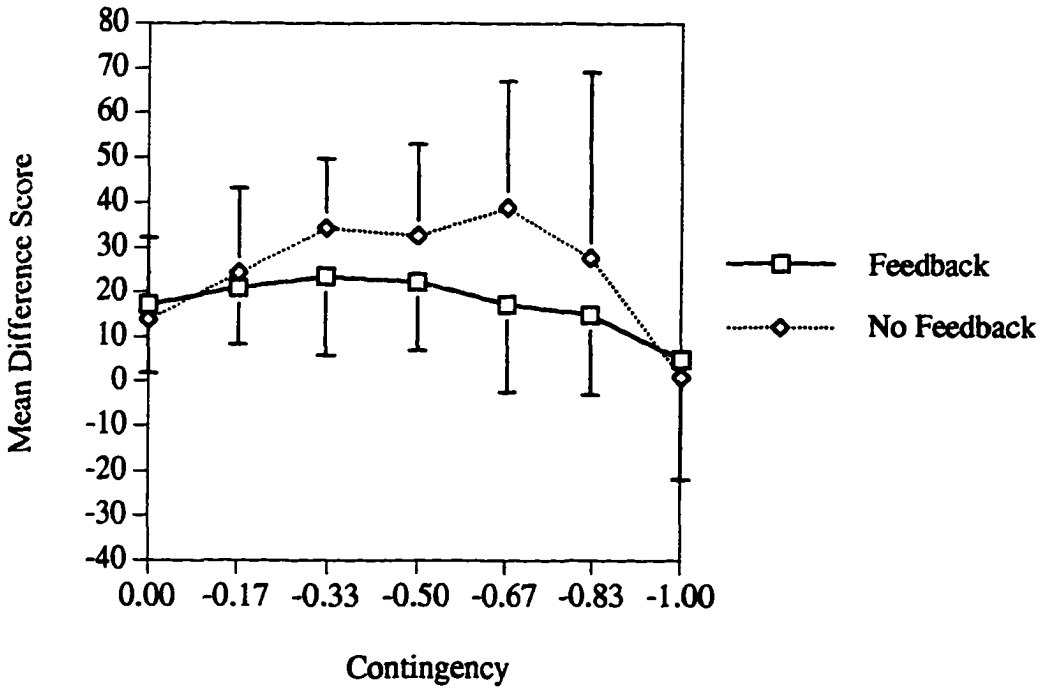
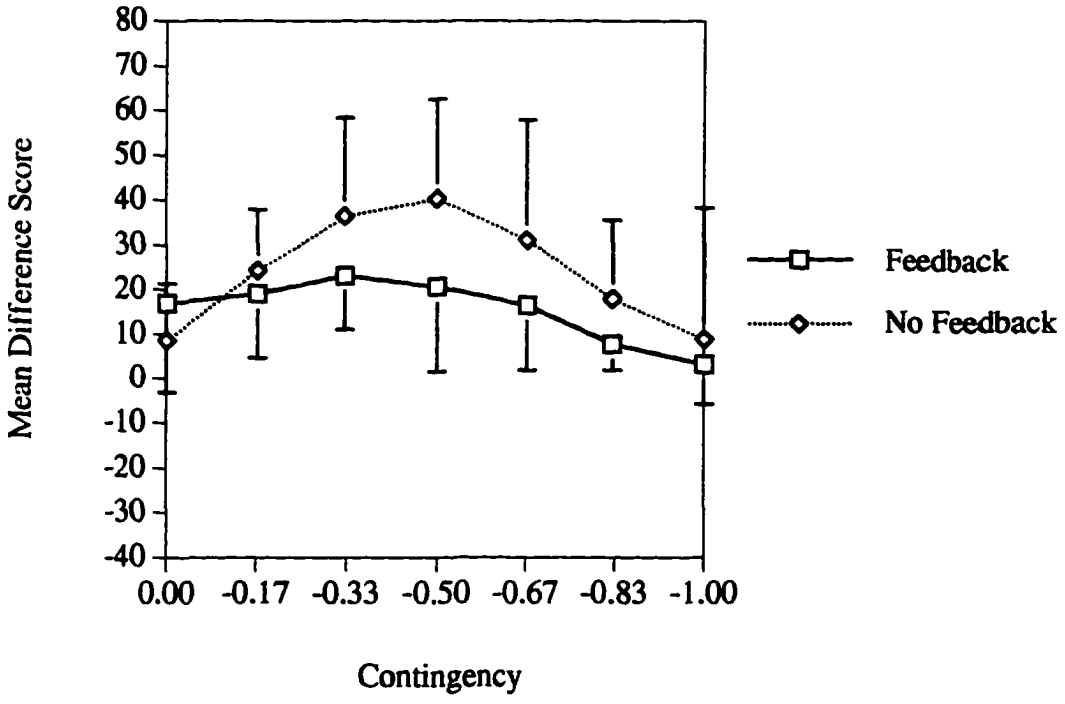


Table 28
Mean Absolute Difference Scores by Level of Contingency for Problem Set 3 (Experiment 2)

| Contingency Group | Condition | | | | | |
|---|---|---|---|---|---|---|
| | Feedback (n = 258) | | | No Feedback (n = 264) | | |
| | n | M | SD | n | M | SD |
| -1.00 | 57 | 3.21 | 8.99 | 58 | 8.86 | 29.43 |
| -.83 | 24 | 7.71 | 5.78 | 27 | 17.89 | 17.74 |
| -.67 | 36 | 16.47 | 14.53 | 42 | 31.10 | 26.78 |
| -.50 | 47 | 20.62 | 19.09 | 37 | 40.32 | 22.29 |
| -.33 | 30 | 23.17 | 12.02 | 34 | 36.44 | 21.99 |
| -.17 | 37 | 19.05 | 14.49 | 39 | 24.31 | 13.52 |
| .00 | 27 | 16.89 | 19.99 | 27 | 8.59 | 12.68 |

Figure 14
Mean Absolute Difference Score by Level of Contingency for Problem Set 3 (Experiment 2)



These data reveal a different pattern from what was observed in Experiment 1. In Experiment 1, the feedback condition had mean difference scores that were relatively consistent across levels of contingency, but were the smallest for contingencies in the .83 and 1.00 groups (contingencies above .75). The no feedback condition had mean difference scores that were larger than those of the feedback condition, but they also were relatively consistent across levels of contingency. The smallest mean difference scores of the no feedback condition were for contingencies in the .83 and 1.00 groups.

In Experiment 2, the feedback condition has a pattern of mean difference scores that is similar to that of the feedback condition in Experiment 1. The mean difference scores were relatively consistent across levels of contingency except for the -.83 and -1.00 groups (problems more contingent than -.75). For these groups, the mean difference scores were the smallest.

The no feedback condition in Experiment 2 did not have consistent mean difference scores across levels of contingency. Instead, the no feedback condition had relatively small mean difference scores for the -1.00 and .83 groups (problems more contingent than -.75) and for the .00 group (contingencies between -.08 and .08). Further, the no feedback condition had relatively high mean difference scores

for the -.67, -.50, -.33 groups (contingencies between -.75 and -.25).

I again suspected that participants in the no feedback condition judged "0" as a default for any problem that was closer to being noncontingent than being perfectly contingent. If this is the case, it would help explain why participants' mean difference scores were relatively small for the .00 contingency group and relatively large for the -.67, -.50, and -.33 contingency groups.

I performed the same analysis for problem set 3 of Experiment 2 that I performed for problem set 3 of Experiment 1. I found the number of times that participants gave a judgment of "0" for contingency problems at each level of contingency. The results for are presented in Table 29.

Table 29
Number of "0" Judgments and Total Judgments for Each Level of Contingency in Problem Set 3 (Experiment 2)

| | Condition | | | | | |
|---|---|---|---|---|---|---|
| | Feedback | | | No Feedback | | |
| Contingency Group | "0"s | Total | % of Total | "0"s | Total | % of Total |
| -1.00 | 0 | 57 | 0 | 1 | 58 | 2 |
| -.83 | 0 | 24 | 0 | 1 | 27 | 4 |
| -.67 | 0 | 36 | 0 | 8 | 42 | 19 |
| -.50 | 2 | 47 | 4 | 15 | 37 | 41 |
| -.33 | 0 | 30 | 0 | 19 | 34 | 56 |
| -.17 | 7 | 37 | 19 | 24 | 39 | 62 |
| .00 | 8 | 27 | 30 | 16 | 27 | 59 |

In the feedback condition, only 30% of the judgments in the .00 contingency group were "0"s. This is a marked contrast to the no feedback condition in which 59% of the judgments for this level of contingency were "0"s. In addition, participants in the no feedback condition judged that over 40% of the problems in the -.17, -.33, and -.50 groups were "0"s.

The use of "0" as a default answer accounts for the participants in the no feedback condition having small mean difference scores for the .00 contingency group and large mean difference scores for the -.67, -.50, and -.33 contingency groups.

Overall, the pattern of mean difference scores in this experiment suggest that there is a difference in discriminability at different levels of contingency. Further, these data suggest that the negative contingencies used in the second experiment are more difficult to discriminate among than the positive contingencies used in Experiment 1.

Response Rate

Wasserman et al. (1983) found a difference in judgmental accuracy as a function of participant's response rate during the experiment. This finding was not supported in Experiment 1 of this dissertation. To assess whether it would be found in Experiment 2, I again classified participants according to mean response rate over the 21 problems of the experiment. The top third of all participants had a mean response rate

above 13.89 (mean response probability = .63). The bottom third of all participants had a mean response rate below 12.85 (mean response probability = .51). The mean difference scores on problem sets 1, 2, and 3 are presented in Table 30 for participants with a high or low response rate.

Table 30
Mean Absolute Difference Scores on the Discrete-Trial Task for Participants with a High or Low Response Rate (Experiment 2)

| | Condition | | | |
|---|---|---|---|---|
| | Feedback (n = 23) [1] | | No Feedback (n = 31) [2] | |
| Problem Set | M | SD | M | SD |
| 1 | | | | |
| High Response Rate | 25.75 | 11.40 | 25.17 | 9.07 |
| Low Response Rate | 21.21 | 12.78 | 20.71 | 10.91 |
| 2 | | | | |
| High Response Rate | 18.40 | 5.74 | 24.60 | 9.21 |
| Low Response Rate | 17.17 | 7.89 | 19.93 | 9.76 |
| 3 | | | | |
| High Response Rate | 15.00 | 5.48 | 23.96 | 8.32 |
| Low Response Rate | 16.38 | 7.49 | 23.61 | 11.86 |

[1] There were 11 participants with a high response rate and 12 participants with a low response rate in the feedback condition.
[2] There were 16 participants with a high response rate and 15 participants with a low response rate in the no feedback condition.

A repeated-measures ANOVA was performed with condition (feedback and no feedback) and response rate (high and low) as between-subject variables. The repeated-measures variable was mean absolute difference scores on the three problem sets. The main effect for response rate was not significant,

$F$ (1,49) = 1.42, $p$ > .10. None of the interactions involving response rate were significant (all $p$ values > .10).

As in Experiment 1, I found no difference in judgmental accuracy as a function of participant's response rate. I again suspect that this is due to the fact that participants in my Experiment 2 were responding at a rate that was closer to an optimal rate of responding than were Wasserman et al.'s (1983) participants. Participants in both the high and low response rate groups in Experiment 2 were within 15 units of an optimal response rate (.63 and .51, respectively).

# GENERAL DISCUSSION

## Principal Findings of Experiments 1 and 2

My primary question in conducting this research was whether feedback and practice would improve participants' judgmental accuracy. The results of Experiment 1 and 2 suggest that feedback combined with practice improved participants' judgmental accuracy on both positive and negative contingency problems. Practice alone, as evidenced by the judgments of participants in the no feedback conditions, did not lead to greater judgmental accuracy.

Experiments 1 and 2 also document the well known psychophysical function for the judgment of contingency. The judgments of participants in the no feedback conditions demonstrated an underestimation of the degree of contingency. As is typical, this underestimation was the greatest for contingencies between -.50 and .50. The judgments of participants in the feedback conditions showed an underestimation of objective contingency on problem set 1, but improved in judgmental accuracy by problem set 3.

## Judgmental Accuracy and Judgment Strategy

Participants' in the feedback conditions became more accurate judges of contingency over the course of the three problem sets. To what should I attribute this improvement? The obvious answer is feedback and practice. But there is also a more fundamental question here. How did these

participants become better judges of contingency? In other words, how do people make contingency judgments?

Considerable research has addressed this question with rule-based analysis (Allan, 1993; Allan & Jenkins, 1983; Wasserman, 1990; Wasserman et al., 1983; Shaklee, 1983; Shaklee & Mims, 1981; Shaklee & Tucker, 1980; Shaklee & Wasserman, 1986). Rule-based analysis is an attempt to identify which of several judgment rules best describe participants' contingency judgments. The assumption behind this type of analysis is that if a particular strategy describes a participants' judgments, the participant may have been using that strategy.

Five rules have been identified as possible judgment strategies (Allan, 1993): conditional probability ($\Delta P$), sum of diagonals ($\Delta D$), frequency of Cell A versus Cell B ($F_{A-B}$), frequency of Cell A versus Cell C ($F_{A-C}$), and frequency of Cell A ($F_A$).

The rule that would lead to accurate judgments in every case is the $\Delta P$ rule (Allan, 1993). $\Delta P$ is the appropriate statistical measure for the relation between two binary variables (Allan, 1980). It represents the probability of an outcome given a response (p[O/R]) minus the probability of an outcome given no response (p[O/no R]). Some research has found that people's judgments of contingency are highly correlated with objective contingency (e.g., Allan & Jenkins, 1980; Alloy & Abramson, 1979; Wasserman et al., 1983;

Wasserman & Shaklee, 1984). This correlation, however, does necessarily mean that people are using the ΔP rule. Under some conditions, the use of other judgment rules will lead to judgments that are consistent with the ΔP rule.

The ΔD rule applies only to 1R/1O contingency problems because it is based on the idea of comparing the number of confirming cases (Cells A and D) with the number of disconfirming cases (Cells B and C): ΔD = (A+D)-(B+C). Use of the ΔD rule promotes judgments of contingency that are perfectly correlated with ΔP when the probability of a response is equal to the probability of no response (p[R] = p[no R], or in terms of a 2 X 2 table, [A+B] = [C+D]).

For participants to use the ΔP and ΔD rules, they must attend to all of the relevant contingency information (all four cells of a 2 X 2 table). Other judgment rules do not require this. The $F_{A-B}$ rule is based on comparing the number of outcomes which occur after a response (Cell A) to the number of responses without an outcome (Cell B). The use of this rule provides judgments of contingency that are perfectly correlated with ΔP when the probability of an outcome is equal to the probability of no outcome (p[O] = .50, or in terms of a 2 X 2 table, [A+C] = [B+D]).

Use of the $F_{A-C}$ rule also requires information from two cells of a 2 X 2 table. This rule is based on comparing the number of outcomes which occur with a response (Cell A) to the number of outcomes which occur without a response (Cell

C). The use of this rule provides judgments that are perfectly correlated with those of $\Delta P$ when the probability of a response is equal to the probability of no response (p[R] = .50, or in terms of a 2 X 2 table, [A+B] = [C+D]).

For participants to use the $F_{A-B}$ and $F_{A-C}$ rules, they need information from two cells of a 2 X 2 table. The use of the $F_A$ rule, in contrast, is based on the number of times that an outcome occurs with a response (Cell A). The $F_A$ rule is often the reported strategy of participants (Smedslund, 1963).

In Experiment 1 of this dissertation, participants were asked about their judgment strategy at the end of each problem set. They responded to the prompt: "Please describe below how you made your judgments on the last seven problems. That is, on what did you base your evaluations?"

Participants' responses were coded by myself and one of my former students who helped collect the data. We classified each response as one of the five rules stated above or "other." We were blind to participants' condition and judgmental accuracy as we coded the data.

Our coding revealed that very few participants clearly state one of the rules as their method of judgment. The vast majority of responses were classified as "other." There were so few participants who stated rules that no analyses were performed on these data.

One of the problems we encountered was that participants lacked a vocabulary to clearly describe the basis for their judgments. Many participants struggled to describe how they had made their judgments. Even when participants had described one of the five judgment rules, they often fumbled for words to state it again after another problem set.

The judgment strategies of some participants suggested that they based their judgments on formal rules. Other subjects did not state formal rules and may have based their judgments on processes they cannot describe. It may be that an open-ended question about judgment strategies asks participants to tell more than they know about the processes of their own judgments (Nisbett & Wilson, 1977). This could be another drawback to the use of open-ended questions.

Shaklee (1983) identified a way to assess participants' rule use without the problems associated with open-ended questions. Shaklee constructed a set of summary table contingency problems in which participants' judgments would be diagnostic of their judgment strategy. Participants who used the $\Delta P$ rule could provide accurate judgments of all the problems, but some of the problems were constructed so that they could also be accurately judged by use of the $F_A$, $F_{A-B}$, or $\Delta D$ rules. Participants' judgment strategies would be inferred from the problems that they judged correctly. Using this type of diagnostic problem set, Shaklee and Wasserman (1986) found that only 3% of their participants showed a

pattern of judgment consistent with the use of the ΔP rule. The most common pattern of judgment among their participants was consistent with the use of the $F_{A-B}$ rule (38% of participants).

How would feedback and practice influence participants' judgments of a diagnostic problem set such as the one used by Shaklee and Wasserman (1986)? This is an empirical question and could be answered with a follow-up experiment. I would predict that feedback and practice would bring participants' judgments in line with the ΔP rule. An experiment of this sort would introduce the use of feedback and practice into the rule-based judgment literature.

## Directions for Future Research

In addition to the above idea for a follow-up experiment, the results of this dissertation suggest additional directions for future research. Here are two ideas for additional experiments.

Jenkins and Ward (1965) found that participants' judgments of contingency were unrelated to the actual contingency of the problems they performed. One criticism of their experimental procedure has been that the 2R/20 task they employed does not promote accurate judgment (Allen and Jenkins, 1980). Can participants make accurate judgments of contingency on a 2R/20 task? Would feedback and practice have the same effect on judgmental accuracy in a 2R/20 task as they do in a 1R/10 task? I suspect that participants can make

accurate judgments on a 2R/20 task and that feedback and practice would have the same effect on judgments to this task as on judgments to a 1R/10 task.

Another experiment would investigate the relation between judgmental accuracy and self-efficacy. In their contrast and assimilation experiment, Clark and Benassi (in press) found that when a judgment task was relatively difficult, participants with high levels of self-efficacy showed more judgmental displacement from an anchor than participants with low levels of self-efficacy. When the judgment task was relatively easy, there was no systematic difference between the amount of judgmental displacement of participants with high versus low levels of self-efficacy. An experiment could be conducted in which one group of participants would judge relatively easy to discriminate contingency problems while another group would judge relatively difficult to discriminate contingency problems (problems between -.50 and .50). What would the effects of feedback and practice be on the self-efficacy of participants in the two groups? How would self-efficacy influence judgmental accuracy in this setting? Feedback and practice might influence self-efficacy, especially for participants who must make difficult judgments. Further, high levels of self-efficacy might lead to greater judgmental accuracy, especially for participants who must make the more difficult judgments.

## Conclusion

Hogarth (1981) identified several ways in which most judgment experiments are different from how we make judgments in the real world. Two of the major differences are that in the real world (a) people receive continual feedback and (b) people make many judgments as part of an ongoing process. In contrast, most experiments provide no feedback on the few problems that participants judge. Hogarth claimed that feedback is not only absent from most experiments, but that its importance is not recognized on a theoretical level. This assertion is true, for example, with respect to Crocker's (1981) review article.

In her article, "Judgment of Covariation by Social Perceivers," Crocker (1981) presented the extant judgment literature as it fit into her conception of the six steps of making judgments in the real world. Her steps are:

> (1) decide what kinds of data to collect, (2) sample cases from the population of cases, (3) interpret the cases (i.e., code the data), (4) recall the data that have been collected and estimate the frequencies of confirming and disconfirming cases, (5) integrate the evidence, and (6) use the estimate as a basis for making predictions or judgments. (p. 273)

Crocker's six steps lead to making "predictions or judgments," but in reading her article it seems that these predictions and judgments are onetime events with no subsequent feedback. One could argue that people begin this

process over again as soon as it is completed, and one could even argue that these steps are used to evaluate feedback. But Crocker's six steps of making judgments are not part of an explicit feedback system. To make it so, a seventh step would need to be added: (7) begin process over again, evaluating feedback that result from judgment.

The contribution of the present research is that it brings judgment of contingency research one step closer to reflecting real-world judgments. As I indicated in the Introduction, sensitivity to covariation is an important prerequisite of adaptive behavior. Luckily, in the real world we usually have feedback and many chances to get it right.

# REFERENCES

Allan, L. G. (1980). A note on measurement of contingency between two binary variables in judgment tasks. Bulletin of the Psychonomic Society, 15, 147-149.

Allan, L. G. (1993). Human contingency judgments: Rule based or associative? Psychological Bulletin, 114, 435-448.

Allan, L. G., & Jenkins, H. M. (1980). The judgment of contingency and the nature of the response. Canadian Journal of Psychology, 34, 1-11.

Allan, L. G. & Jenkins, H. M. (1983). The effect of representations of binary variables on judgment of influence. Learning and Motivation, 14, 381-405.

Alloy L. B., & Abramson, L. Y. (1979). Judgment of contingency in depressed and nondepressed students: Sadder but wiser? Journal of Experimental Psychology: General, 108, 441-485.

Alloy, L. B., & Tabachnik, M. (1984). Assessment of covariation by humans and animals: The joint influence of prior expectations and current situational information. Psychological Review, 91, 112-149.

Bandura, A. (1986). Social foundations of thought and action: A social cognitive theory. Englewood Cliffs, NJ: Prentice-Hall.

Bandura, A. (1990). Conclusion: Reflections on nonability determinants of competence. In R. J. Sternberg & J. Kolligian, Jr. (Eds.), Competence considered (315-362). New Haven, CT: Yale University Press.

Bandura, A., & Wood, R. (1989). Effect of perceived controllability and performance standards on self-regulation of complex decision making. Journal of Personality and Social Psychology, 56, 805-814.

Benassi, V. A., & Mahler, H. I. M. (1985). Contingency judgments by depressed college students: Sadder but not always wiser. Journal of Personality and Social Psychology, 49, 1323-1329.

Bobko, P., & Karren, R. (1979). The perception of Pearson product moment correlations from bivariate scatterplots. Personnel Psychology, 32, 313-325.

Clark, S. C. & Benassi, V. A. (in press). Judgment of contingency: Contrast and assimilation, displacement of judgments, and self-efficacy. Social Behavior and Personality: An International Journal.

Crocker, J. (1981). Judgment of covariation by social perceivers. Psychological Bulletin, 90, 272-292.

Herrnstein, R. J. (1966). Superstition: A corollary of the principles of operant conditioning. In W. K. Honig (Ed.) Operant behavior: Areas of research and application. New York: Appleton-Century-Crofts

Hogarth, R. M. (1981). Beyond discrete biases: Functional and dysfunctional aspects of judgmental heuristics. Psychological Bulletin, 90, 197-217.

Jenkins, H. M., & Ward, W. C. (1965). Judgment of contingency between responses and outcomes. Psychological Monographs, 79(1, Whole No. 594).

McFarland, D. J. (1971). Feedback mechanisms in animal behavior. London: Academic Press.

Newman, S. E. & Benassi, V. A. (1989). Putting judgments of control into context: Contrast effects. Journal of Personality and Social Psychology, 56, 876-889.

Nisbett, R. E., & Wilson, T. D. (1977). Telling more than we can know: Verbal reports on mental processes. Psychological Review, 84, 231-259.

Peterson, C. (1980). Recognition of noncontingency. Journal of Personality and Social Psychology, 38, 727-734.

Shaklee, H. (1983). Human covariation judgment: Accuracy and strategy. Learning and Motivation, 14, 433-448.

Shaklee, H., & Mims, M. (1981). Development of rule use in judgments of covariation between events. Child Development, 52, 317-325.

Shaklee, H., & Tucker, D. (1980). A rule analysis of judgments of covariation between events. Memory & Cognition, 8, 459-467.

Shaklee, H., & Wasserman, E. A. (1986). Judging interevent contingencies: Being right for the wrong reasons. Bulletin of the Psychonomic Society, 24, 91-94.

Sherif, M., Taub, D., & Hovland, C. I. (1958). Assimilation and contrast effects of anchoring stimuli on judgments. Journal of Experimental Psychology, 55, 150-155.

Smedslund, J. (1963). The concept of correlation in adults. Scandinavian Journal of Psychology, 4, 165-173.

Ward, W. C., & Jenkins, H. M. (1965). The display of information and the judgment of contingency. Canadian Journal of Psychology, 19, 231-241.

Wasserman, E. A. (1990). Detecting response-outcome relations: Toward an understanding of the causal texture of the environment. In G. H. Bower (Ed.), The psychology of learning and motivation (Vol. 26, pp. 27-82) San Diego, CA: Academic Press.

Wasserman, E. A., Chatlosh, D. L., & Neunaber, D. J. (1983). Perception of causal relations in humans: Factors affecting judgments of response-outcome contingencies under free-operant procedures. Learning and Motivation, 14, 406-432.

Wasserman, E. A., & Shaklee, H. (1984). Judging response-outcome relations: The role of response-outcome contingency, outcome probability, and method of information presentation. Memory & Cognition, 12, 270-286.

Well, A. D., Boyce, S. J., Morris, R. K., Shinjo, M., & Chumbley, J. I. (1988). Prediction and judgment as indicators of sensitivity to covariation of continuous variables. Memory and Cognition, 16, 271-280.